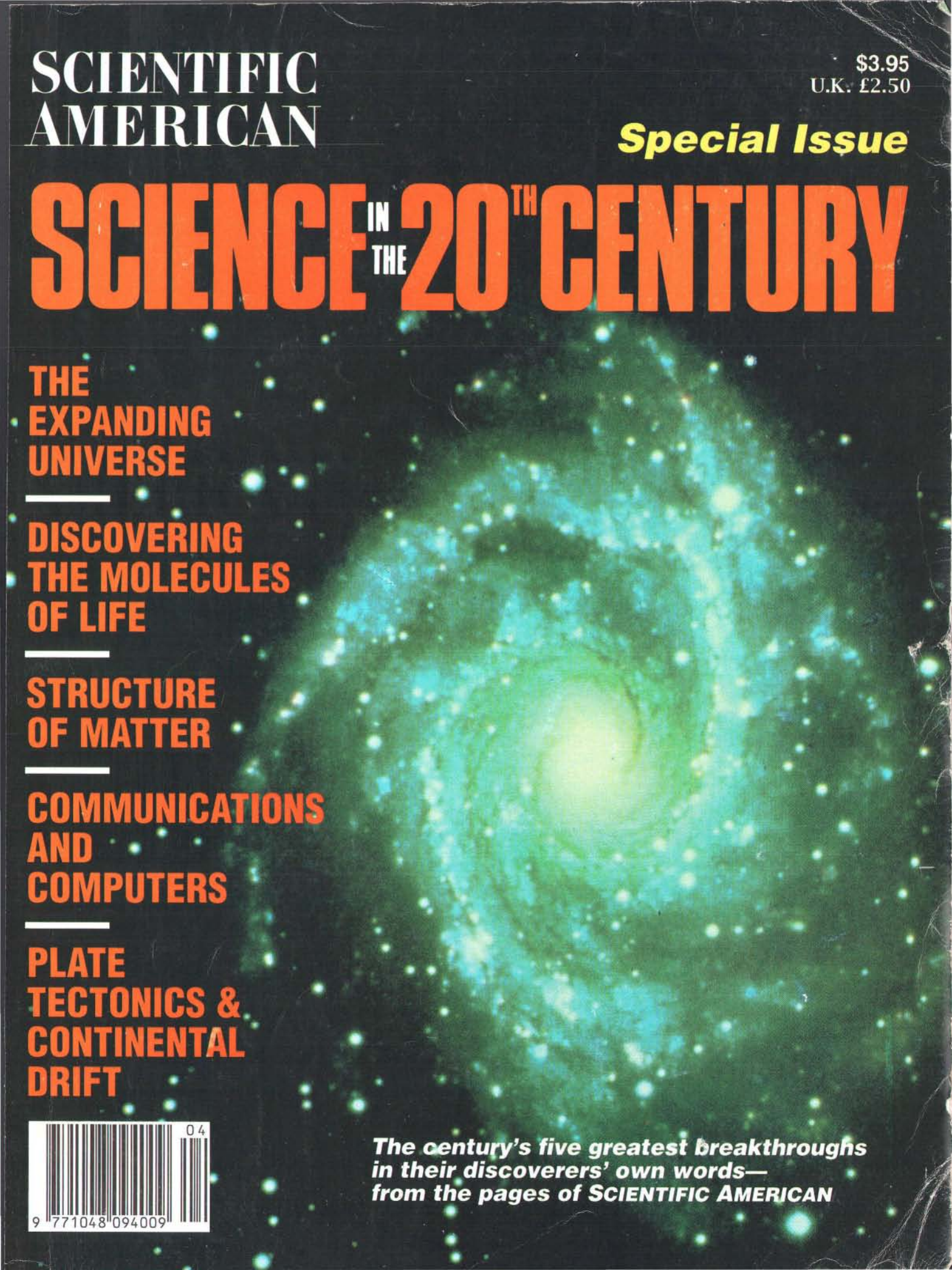
SCIENTIFIC AMERICAN

$3.95
U.K. £2.50

Special Issue

SCIENCE IN THE 20TH CENTURY



THE EXPANDING UNIVERSE

DISCOVERING THE MOLECULES OF LIFE

STRUCTURE OF MATTER

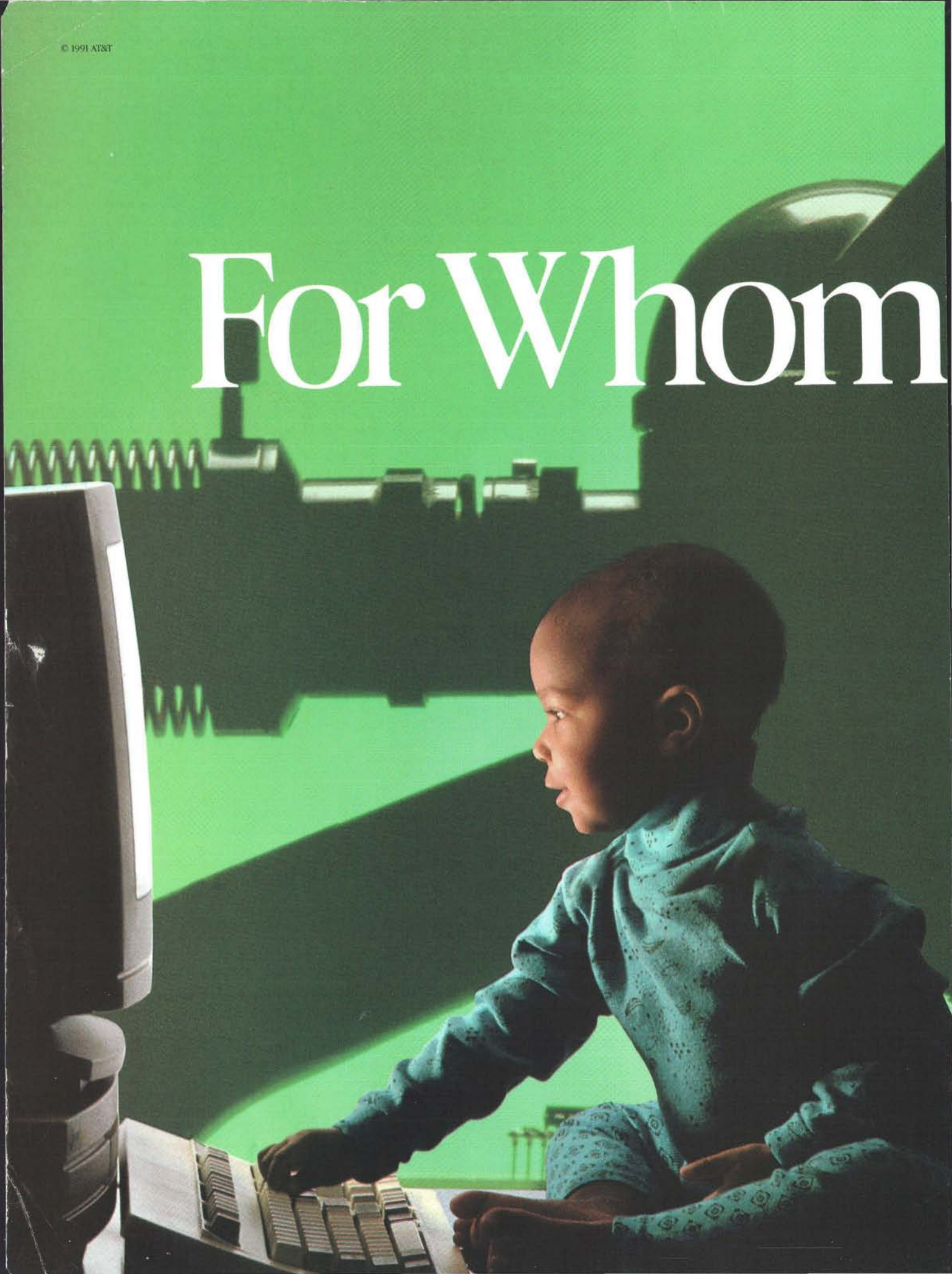
COMMUNICATIONS AND COMPUTERS

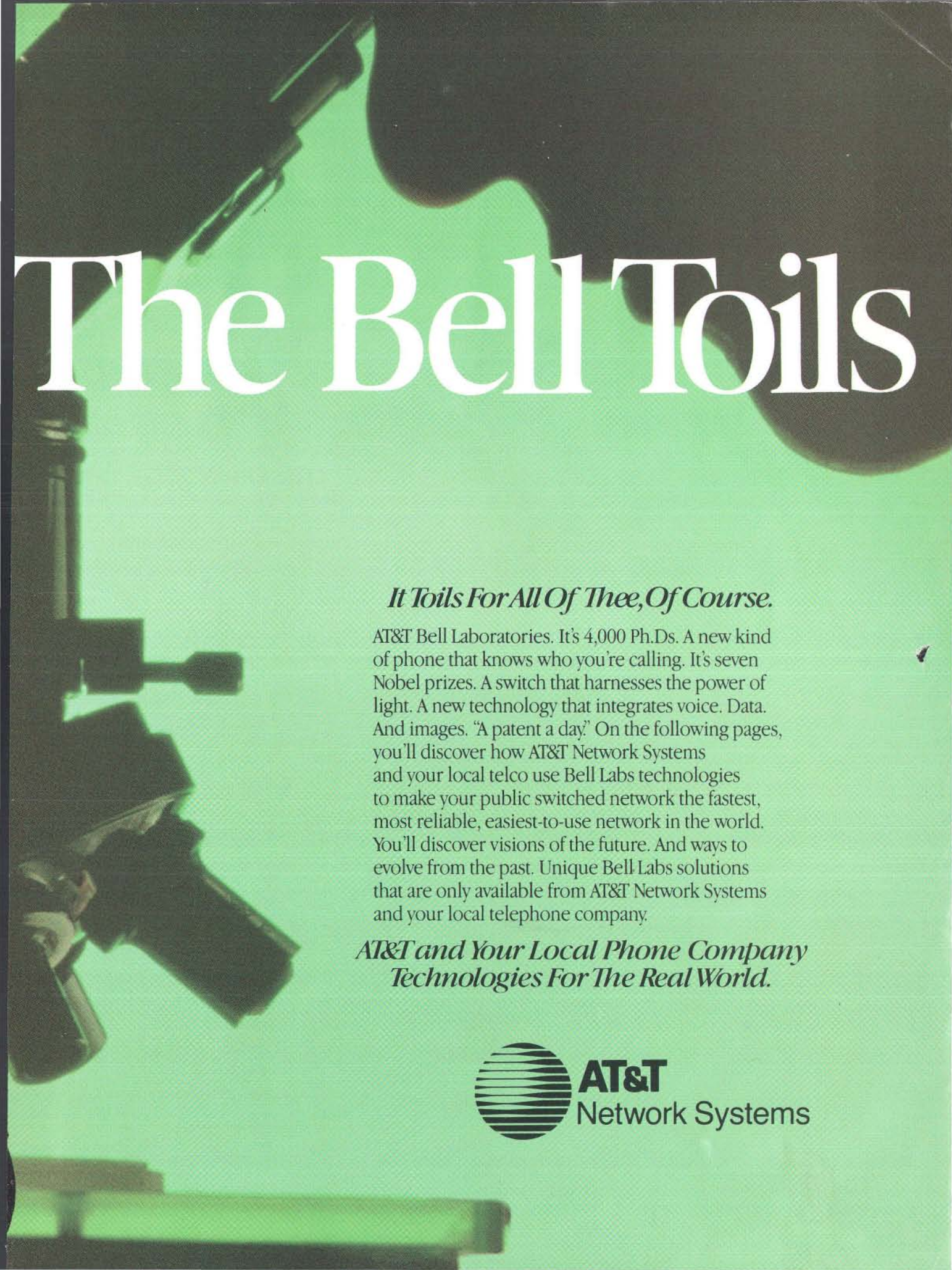
PLATE TECTONICS & CONTINENTAL DRIFT

9 771048 094009 04

The century's five greatest breakthroughs in their discoverers' own words— from the pages of SCIENTIFIC AMERICAN

For Whom





It Toils For All Of Thee, Of Course.

AT&T Bell Laboratories. It's 4,000 Ph.Ds. A new kind of phone that knows who you're calling. It's seven Nobel prizes. A switch that harnesses the power of light. A new technology that integrates voice. Data. And images. "A patent a day." On the following pages, you'll discover how AT&T Network Systems and your local telco use Bell Labs technologies to make your public switched network the fastest, most reliable, easiest-to-use network in the world. You'll discover visions of the future. And ways to evolve from the past. Unique Bell Labs solutions that are only available from AT&T Network Systems and your local telephone company.

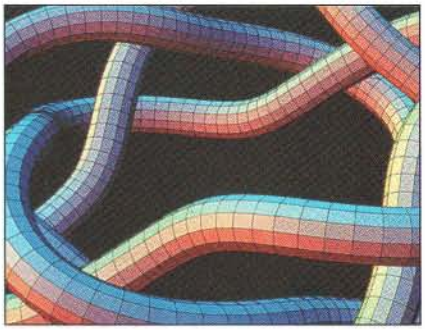
AT&T and Your Local Phone Company
Technologies For The Real World.

AT&T
Network Systems

SCIENTIFIC AMERICAN

SCIENCE IN THE 20TH CENTURY

-1 CHARGE	0 CHARGE
ELECTRON MASS: ABOUT 5.11 X 10^{-4} GeV	**ELECTRON NEUTRINO** MASS: <2 X 10^{-8} GeV
MUON MASS: 0.106 GeV	**MUON NEUTRINO** MASS: <2 X 10^{-4} GeV
TAU MASS: 1.78 GeV	**TAU NEUTRINO** MASS: <0.035 GeV



The millennial turn lies just 10 years ahead. Yet the century has already established itself as a watershed in the history of our species' effort to understand and master the forces of nature. For the first time, we have discerned the fundamental character of matter, begun to comprehend the structure of the universe and gained direct understanding of the chemical process that we call life. In this special issue, key researchers tell this story of achievement in their own words, as they appeared in the pages of SCIENTIFIC AMERICAN. Although the record presents accomplishment, it also offers glimpses of the experience of doing science. The shoulders on which the great scientists of the 21st century will stand are very human ones.

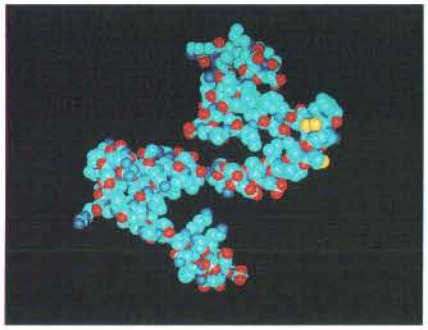
Jonathan Piel
Editor

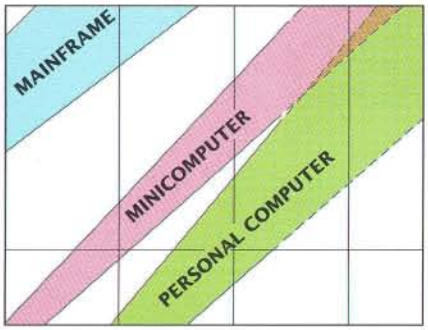
THE STRUCTURE OF MATTER

THE EXPANDING UNIVERSE

SCIENTIFIC AMERICAN *Science in the 20th Century* (ISSN 1048-0943), Special Issue Vol. 3, Number 1, 1991, published by Scientific American, Inc., 415 Madison Ave., N.Y., N.Y. 10017. To purchase additional quantities: 1–9 copies: US$3.95 each plus US$1 (US$2 foreign) per copy postage and handling, 10 or more copies: US$3.95 each, postpaid. Send payment to: Scientific American, Dept. 20C, 415 Madison Ave., N.Y., N.Y. 10017. Canadian G.S.T. #R127387652.

Special Issue/1991





SCIENTIFIC AMERICAN

Science in the 20th Century

Humankind's "passion for comprehension" has given our species unmatched understanding of and control over the natural universe

by Jonathan Piel

"There exists a passion for comprehension," Albert Einstein wrote in the pages of SCIENTIFIC AMERICAN 41 years ago. "Without this passion, there would be neither mathematics nor natural science."

This special issue has been assembled to celebrate that passion's crowning achievement: the powerful body of knowledge and technology that is SCIENCE IN THE 20th CENTURY. To tell the story, my colleagues on the Editorial Board and I have chosen to present five distinct adventures in exploration that are central to the broad array of accomplishment in the sciences that our century has witnessed. They appear under the following rubrics: The Nature of Matter, The Expanding Universe, Discovering the Molecules of Life, The Revolution in Geophysics and The Physics of Computing and Communications. A set of articles from the pages of SCIENTIFIC AMERICAN embodies the themes. The first in each group is historic—it represents ground-breaking work by the scientist responsible for the achievement (Einstein, Watson and Crick) or someone close to the achievers. The other articles represent milestones along the way to the current state of knowledge.

Could such a celebration be premature? After all, the century has another decade to run. True, but were the century to end tomorrow, the scientific achievements it has seen would still stand out strongly against the historical background.

For the first time, our species has discerned the fundamental nature of physical and biological reality. We know that matter consists of a handful (although a large handful) of fundamental particles. Interacting through subtle forces, they form a universe that apparently had an awesome beginning. The cosmos now consists of an array of galactic superclusters, punctuated by great voids, across which shine the energy of pulsars, supernovas and ordinary stars, at least one of which supports a life-bearing planet. Equally exciting, we now understand the life that has elaborated itself from atmosphere, hydrosphere and geosphere.

Supreme knowledge confers supreme power. The 20th century is the first in which inquiry into nature has become the source of technologies that amplify the power of the human mind and may even render human intelligence redundant. The 20th century is also the first in which technology (augmented by the growth in the human population that it has made possible) has become a natural force—one that threatens the planet's habitability.

The 20th century has experienced a sea change in the relationship between science and technology. In the 18th century the members of Birmingham's Lunar Society learned more about physics from their mill machinery than physics contributed to productivity. The study of illness has traditionally illuminated biological processes; now medicine takes lessons from molecular biology.

Finally, science has become a global profession with its own rules and culture. It is a profession institutionalized in university, government and industrial research laboratories. Political leaders and entrepreneurs have begun to listen to what science has to say.

These intellectual and cultural accomplishments grow from varied and ancient rootstock. The most fertile 20th-century offshoot consists of two elegant, counterintuitive bodies of theory. One is quantum theory, the work of such great physicists as Max Planck, Niels Bohr, Max Born, Erwin Schrödinger, Wolfgang Pauli, P.A.M. Dirac, Werner Heisenberg and Louis de Broglie. The other theory is relativity, primarily the work of Einstein. Einstein's theory has enabled us to think about the birth and ultimate fate of the universe. Quantum theory has illuminated the nature of matter, yielding powerful technologies in electronics, energy and materials.

Both theories owe their existence to the faith of physicists in nature's underlying simplicity, a faith that has persisted—sometimes in the face of adversity—for almost three millennia (Democritus's proposal, made in 400 B.C., that matter must ultimately consist of identical, indivisible entities, atoms, provides a reasonable benchmark.) At the turn of this century, the work of J. J. Thomson, Ernest Rutherford and others revealed that the atom is not truly fundamental, that it has structure. The atom could be viewed as a tiny solar system: the positively charged nucleus orbited by negative electrons.

It was this tiny solar system model that Bohr, Born and others so deftly reformulated in quantum mechanical terms as a nucleus surrounded by electrons, each held at its own well-defined energy level, forming a "probability" cloud. As he considered the implications of quantum theory and relativity, Dirac observed that the electron should have a positively charged counterpart. That prediction was subsequently confirmed (in 1932) when C. D. Anderson discovered the positron. Physics had stepped through the looking glass into the world of antimatter.

With James Chadwick's discovery of the chargeless neutron—also in 1932—

My thanks to James T. Rogers, an old friend and colleague, for invaluable help in preparing this introductory article.

the roster of atomic components seemed both short and complete: it comprised the proton, the neutron (which decays into a proton, electron and neutrino) and the electron itself. Of the three, the proton and electron seemed truly fundamental—that is, lacking further structure.

Even as this scheme took shape, it began, in Alice's words, to grow "curiouser and curiouser." Could these fundamental particles, both matter and antimatter, have structure too? Were they truly unitary and indivisible? Accelerator and detector technology began to provide an answer. Devices such as the cyclotron, the Cockcroft-Walton accelerator and the Van de Graaf generator were elaborated after World War II into even more sophisticated instruments. As the energy of the collisions that these machines could stage reached into millions and then billions of electron volts, the neat picture of a few fundamental particles began to dissolve. At Brookhaven National Laboratory, the European laboratory for particle physics (CERN) and the Stanford Linear Accelerator Center (SLAC) and elsewhere, hundreds of subnuclear entities were observed in bubble chambers, spark chambers and other devices.

If the physicists could not achieve simplicity, they could at least strive for order. In 1961 Murray Gell-Mann of the California Institute of Technology and Yuval Ne'eman of the University of Tel Aviv independently created a symmetry theory that grouped the particles into families according to charge, spin and other properties. Gell-Mann and George Zweig even provided an insight that promised some restoration of simplicity and unity. They proposed that the proton and other hadrons consist of fractionally charged particles, which Gell-Mann called quarks. (When he named the quark, Gell-Mann also afflicted physics with a penchant for whimsy. Every worker seems compelled to crown his or her discovery with a tortured moniker.)

This line of work culminated during the 1970s in the theory now called the Standard Model of elementary particles. According to the Standard Model, particles that compose structure are classed as fermions and those that serve as the messengers of the forces between them are bosons. Fermions can be organized into families consisting of two quarks and two leptons (one of which is a neutrino) [*see illustration on next page*]. Four forces enable the particles to interact: the electromagnetic force, mediated by the photon; the strong force, carried by the gluon, which binds the nuclear constituents; the weak force, carried by *W* and *Z* bosons, which mediates some radioactive processes as well as neutrino interactions; and gravity, mediated by a presumed but undetected graviton.

Was simplicity truly secure? In 1974 the psi/J particle was detected at SLAC and at Brookhaven. Together with Nobel Prize–winning work done at SLAC some three years previously by Henry W. Kendall, Jerome Friedman and Richard E. Taylor demonstrating that the proton has structure, the later result confirmed the existence of the quark. But might not there be many families of quarks and their related leptons? In 1990 at CERN, workers using the LEP Collider reported measurements of the decay of the *Z* boson, which demonstrated that there are indeed only

GREAT GALAXY in Andromeda symbolizes a major achievement of science in the 20th century: deepening insight into the structure of the universe.

How Fundamental Particles Form Three Families Of Matter

	PARTICLES			
	LEPTONS		QUARKS	
FAMILY	-1 CHARGE	0 CHARGE	-1/3 CHARGE	2/3 CHARGE
ELECTRON	ELECTRON MASS: ABOUT 5.11 X 10^{-4} GeV	ELECTRON NEUTRINO MASS: <2 X 10^{-8} GeV	DOWN MASS: ABOUT 0.01 GeV	UP MASS: ABOUT 0.01 GeV
MUON	MUON MASS: 0.106 GeV	MUON NEUTRINO MASS: <2 X 10^{-4} GeV	STRANGE MASS: ABOUT 0.15 GeV	CHARM MASS: ABOUT 0.01 GeV
TAU	TAU MASS: 1.78 GeV	TAU NEUTRINO MASS: <0.035 GeV	BOTTOM MASS: ABOUT 5.5 GeV	TOP (UNOBSERVED) MASS: >89 GeV

THE FAMILIES OF MATTER number three, each consisting of two quarks and two leptons. The electron family provides the constituents of ordinary matter (up and down quarks combine to form protons and neutrons).

three families of fundamental particles.

Still, as Jack Steinberger of the CERN group says, the physicists cannot go home. Many elements of the Standard Model must be assumed—they do not emerge naturally. A notable example is mass, thought to be mediated by a particle known as the Higgs boson. Theorists also work to extend the unification of the electromagnetic and weak force to include the strong force and perhaps even gravity, first leading to a grand unified theory (GUT) and then to a TOE, the theory of everything.

Creative tension between experiment and theory inspires work in other areas of physics as well. Investigators strive to explain such phenomena as high-temperature superconductivity, quasicrystals, the behavior of magnetically confined plasmas and chaotic systems.

While some physicists look inward toward reality on the scale of a proton (10^{-13} centimeter), astronomers, extending their senses through optical telescopes, radio telescopes, space probes and computers, try to encompass billions of light-years of space. In the 1920s Edwin P. Hubble observed variable stars in several galaxies and in that way established the distance of the galaxies from Earth. Exploring the galaxies further, he found that the light reaching Earth from them was shifted toward the red end of the electromagnetic spectrum. The redshift, the result of the Doppler effect, meant the galaxies were moving away from Earth. (The rate of speed increases with distance.)

Hubble's work and its elaboration by other astronomers undermined the idea of a static universe that serenely existed without beginning and without end. Utilizing Einstein's general theory of relativity to make sense of Hubble's observations, the Russian physicist Alexander Friedmann proposed that the expanding universe might meet one of three fates, depending on the density of matter and energy and its effect on the curvature of space. The universe, Friedmann posited, might be so strongly curved that it would ultimately fall back in on itself, it might be negatively curved and thus expand forever, or it might be flat, expanding forever at a decreasing rate. It was Roger Penrose and Stephen W. Hawking in 1970 who demonstrated that Einstein's theory of general relativity did not merely point to but required a big bang origin.

Radio astronomy, a uniquely 20th-century invention, provided additional support for this picture of the universe's origin. In the 1960s Arno A. Penzias and Robert W. Wilson of Bell Telephone Laboratories discovered the cosmic background radiation, a uniform microwave glow suffusing the entire sky. The idea that the initial cataclysm would leave such a faded remnant had been developed over the years by George Gamow, the Russian-American physicist, and by Robert Dicke and James Peebles of Princeton University. Penzias and Wilson won the Nobel Prize for Physics in 1978.

Another line of supporting evidence comes from nuclear physics. Observation of the cosmic abundance of helium 4 and other light elements agrees well with big bang predictions. Comparison of theory and observation requires that there are three or at the most four families of fundamental particles, in agreement with the number determined from particle physics experiments.

Despite such success, the big bang model seems to encounter a steady flow of observational challenge. First, there does not appear to be enough visible matter to account for the strength of the gravitational force at work in the universe. Like resourceful chess masters improvising a defense, some theorists have proposed that 90 percent or more of the matter in the universe is cold, dark and therefore invisible.

Then there is the horizon problem. In the standard big bang model, there was never enough time for what is now the visible universe to have reached thermal equilibrium. Therefore, it is difficult to explain why the cosmic background radiation appears to be the same temperature in all directions. The right conditions must be assumed; they do not emerge from the theory.

The theory has also to address the smoothness problem. The primordial gas must have contained fluctuations in density that served as seeds for the galaxies and galactic structures of the existing universe. Yet the inhomogeneities present in a normal gas would be far too extreme. Again, instead of explaining, the theory must assume the appropriate degree of inhomogeneity.

Finally, the flatness problem looms. Why does the visible universe seem to have just the curvature that suggests a gently decelerating rate of expansion?

The inflationary universe model, described in this issue by Alan H. Guth of M.I.T. and Paul J. Steinhardt of the University of Pennsylvania, addresses these problems with a high degree of success. Their model is a critical refinement of the big bang–cold dark matter model. According to it, the currently observed universe evolved from a minute region, embedded in a much larger universe, which began to expand suddenly and dramatically at 10^{-35} second after the big bang.

The inflationary model dispatches the horizon problem: its assumptions predict that our observable neighborhood began as a bud of space-time small enough to have reached thermal equilibrium before inflation began. The smoothness problem? The inflationary model generously offers two possible resolutions: quantum effects acting at the end of inflation can produce density fluctuations, or the sudden end of inflation can capture—the way rapidly forming ice traps air bubbles—inhomogeneities sufficient to seed the formation of galactic structures. Flatness? Imagine a large balloon being rapidly inflated. Any restricted region on its surface will appear relatively flat.

Yet the game has not ended. The discovery of vast structures of galaxies from optical redshift surveys and from the spectacular yield of the *Infrared Astronomical Satellite* needs to be rec-

onciled with the cold dark matter–inflation model.

Applied to biology, modern physics has catalyzed another great cascade of understanding that marks the 20th century's achievements in science: knowledge of the molecules of life. The ability to comprehend biology as interaction between molecules enables investigators to uncover fundamental laws governing genetics, development, nervous system function, metabolism and other processes.

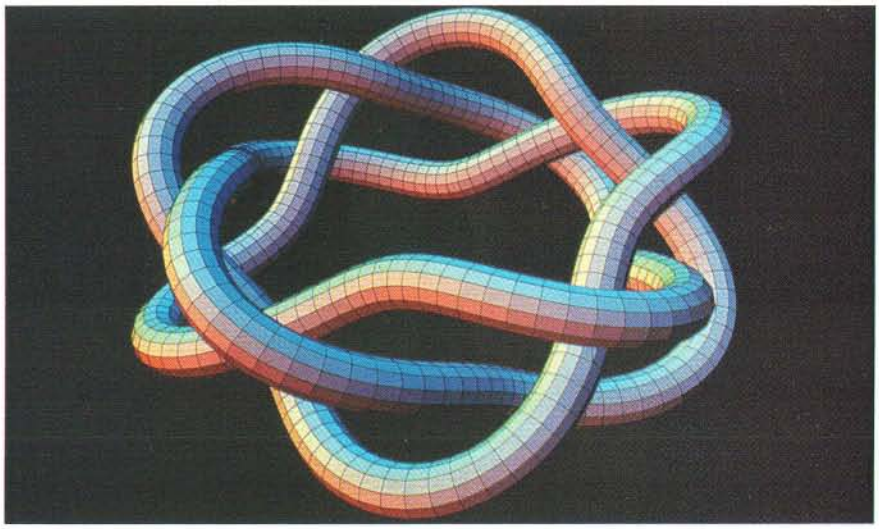
What emerges from their work is a system of genetic controls, metabolic cycles and signaling systems so robust and adaptable that it has elaborated itself into millions, perhaps 10 million or more, separate species. In mice and men as well as in newt and bat, chemical messengers dock with protein receptors protruding from the undulating surfaces of cell membranes. Ions flow through molecular valves. Antibodies mark foreign materials in the body for destruction.

Although two kinds of nucleic acid, DNA and RNA, were known in 1920, it was not until 1944 that Oswald Avery and his associates at the Rockefeller Institute found that DNA is the carrier of genetic information. By 1953 James D. Watson and F. H. C. Crick had worked out the structure of DNA.

They found the double helix, now so familiar as to have taken its place with the Egyptian sandal strap as a symbol of life. Knowledge of DNA's structure provided the key to comprehending its function: to perpetuate genetic information and guide the making of protein molecules, which constitute the cell's physical structure and metabolic machinery. RNA serves as an intermediary, conveying the instructions carried in the DNA to the cellular factories involved in assembling proteins.

Determining the three-dimensional structure of a protein was once a formidable task. Today advances in crystallographic techniques and the computerization of data have created a flood of protein structures. Indeed, the rate of production swamps the ability of biologists to relate structure to function.

Another major preoccupation is deciphering the rules according to which proteins fold. All proteins begin as long linear chains. Each assumes a particular shape, determined by the constituent amino acids. A protein becomes globular or rodlike or takes some other form in a matter of seconds—far too short a time to run through all the possible coilings and foldings. Several workers have proposed various models, some of which assume that a protein compacts into a preliminary shape that restricts the number of pathways it can follow to its final configuration.



KNOT THEORY, represented here by a closed three braid, is a branch of mathematics that—pursued for its own sake—is generating progress in other fields. It has begun to illuminate such diverse areas as molecular biology and theoretical physics. The computer enhances exploration of such complicated constructions.

What emerges from all the knowledge that has been amassed are some of the most exciting questions that can be asked about living beings.

How are genes turned on and off? Workers in many laboratories have identified sites on DNA called promoters and enhancers that appear to moderate gene expression. Some of these genes regulate cell growth. When these systems fail, they can precipitate the development of cancer. Indeed, a whole class of growth-regulating genes, called oncogenes, commands intense scrutiny.

How do complex organisms develop from single cells? How, in other words, do cells specialize? While all cells in the body carry a basic common machinery, each of the hundreds of cell types fulfills a unique function in one or another tissue. The genetic information in all nonreproductive cells is identical. Consequently, each cell performs a specific job that engages only part of its genetic information. How are certain genes potentiated while others are kept dormant? How do cells in an organ such as the liver know that they are to function as liver cells and not as those of muscle? Exciting answers are emerging.

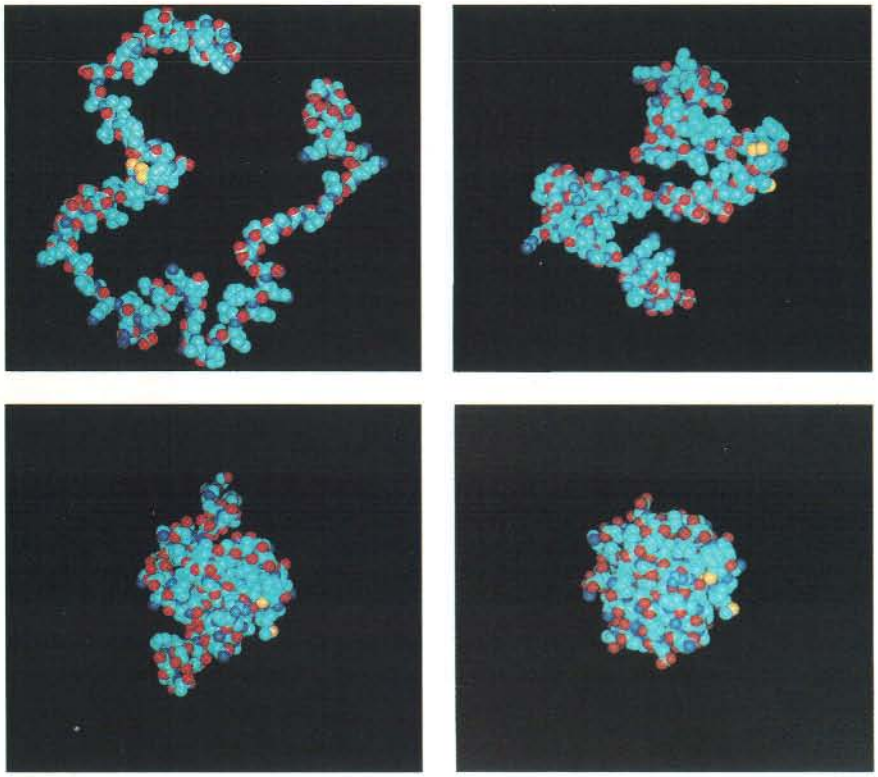
How does a human being, a fish or a fruit fly grow from a fertilized egg to a viable adult individual? A system of genes that communicate through complicated signaling hierarchies acts as a program that creates shape and form. How do cells within the growing tissue join? How do they know where they are? Cellular adhesion molecules appear, with the timing of a trapeze artist, at just the right moment in embryonic development. They enable a cell to identify and adhere to attachment sites.

The burgeoning understanding of molecular biology has also illuminated immunity and the immune system. Charged with distinguishing self from nonself, the immune system consists of an orchestrated array of specialized cells, each kind responsible for a specific task: identifying, marking, destroying and eliminating microorganisms, toxins and other threats.

To identify and precipitate the destruction of foreign material, a cell type called the *B* lymphocyte manufactures antibodies in millions of varieties, each specific to a particular antigen. Both invading agents and the body's own cells display antigens. In a normal individual, antibodies respond exclusively to foreign matter, ignoring the host's cells.

How is the host protected from the immune system? Workers are testing two hypotheses. One holds that cells making antibodies against the host are destroyed early in life; the other that they are merely suppressed. Whatever the mechanism, malfunction can produce a spectrum of autoimmune diseases, lead to tolerance of malignancy or invite devastating infections.

As the fine structure of the system emerges, new ways to fight a wide range of illnesses seem likely. That prospect has become critical. Although old diseases are cured, new ones such as AIDS, Lyme disease or Legionnaire's disease break out, threatening the balance between ourselves and the organisms that prey on us. Immunology of-



FOLDING OF A PROTEIN can be modeled only approximately because investigators still grapple with a biological puzzle: What forces determine how a newly made protein winds into the specific shape that enables it to perform a crucial task in a living cell? Here the protein thioredoxin, initially an open and unstable chain (*a*), becomes increasingly compact (*b* and *c*), ultimately adopting a spherical shape (*d*). The intermediate structures are hypothetical because the actual ones are not fully known. In these models of the process, white represents carbon; red, oxygen; blue, nitrogen; and yellow, sulfur.

fers the only hope of dynamic defense against this protean hazard.

Imagine the situation that would prevail had the AIDS pandemic struck just 20 years ago. The precise mechanism of action of HIV, which attacks a subpopulation of immune cells, the *T4* lymphocytes, would have been far beyond the reach of clinical knowledge; so would have many if not all of the possible strategies for controlling and perhaps curing the disease.

Finally, biologists have begun to probe the grandest question of all: How does the human brain work? What is the chemical and physiological basis of memory and the ability to associate impressions and thoughts? What is the relation between brain and mind?

In order to explore the molecular biology of living beings, investigators have created a subtle and diverse kit of tools: labeled antibodies, monoclonal antibodies, recombinant DNA, restriction fragment length polymorphisms (RFLPs), the polymerase chain reaction. These and other methods constitute biotechnology: the ability to create safer vaccines, bolster or even replace the immune system, maintain genetic function and combat cancer. As Robert A. Weinberg of M.I.T. says, "We are no longer the victims of biology but the masters of it."

The question of how the surface of the earth assumed its form has always engaged us. But no one came close to finding a unifying theory until Alfred Wegener, a German meteorologist, put forward in 1912 the concept of continental drift. The theory holds that the major landmasses have reached their present positions by drifting apart after the breakup some 200 million years ago of a single giant continent, which he called Pangaea.

Not until after World War II did the evidence begin to accumulate, largely as a result of studies financed by the U.S. Navy in the interest of learning more about the milieu in which its submarines and surface ships functioned. Sonar images showed marine geologists Bruce Heezen and Marie Tharp of Columbia's Lamont-Doherty Geological Observatory a deeply rifted ridge on the seafloor in the mid-Atlantic. Similar mid-ocean ridges were discovered in the Pacific and other oceans. The rifts contain basalt of recent origin.

Additional evidence from cores taken by the Deep-Sea Drilling Project showed that the seafloor rocks nearest the ridges were newer than the ones farther away and that all of them were much younger than most of the rocks found on the continents. This and other evidence indicated that molten rock wells up from the mid-ocean ridges, forming new ocean floor as it cools. The new floor moves away from the rift and either helps ocean basins gradually grow larger (the modern Atlantic exemplifies the process) or else plunges downward in deep ocean trenches near the continents, returning to the mantle, as it does at the margins of the Pacific.

Such considerations led to the development of the plate tectonic theory by a wide array of earth scientists, including Harry Hess and Robert Dietz in the U.S., J. Tuzo Wilson in Canada, and Edward Bullard, F. J. Vine and D. H. Matthews in England. According to this theory, the outer regions of the earth consist of the lithosphere—a rigid outer shell ranging in thickness from 50 to 150 kilometers that encompasses both the outermost crust and the upper part of the mantle. Below lies a weaker and hotter part of the mantle called the asthenosphere. The mid-ocean ridges are one of several kinds of boundaries that divide the lithosphere into seven major plates and several minor ones. Carrying the continents, the plates move about on the earth's surface, rather like rafts, riding on the plastic asthenosphere. Earth scientists still puzzle over the precise nature of the mechanism that drives the plates. Hot convecting plumes deep in the mantle are the leading candidate.

In addition to the flows of magma from mid-ocean rifts and volcanoes, there are stationary "hot spots." Like a welding torch on an assembly line, the hot spot remains stationary, whereas the plate moves. Thus, the hot spot can build a chain of volcanic islands, such as the Hawaiian group.

Some workers speculate that the ponderous waltz of the earth's crustal fragments may have affected the climate of our planet. Both the North American continent and the Eurasian continent bear dramatically raised plateaus, the result of crustal collision. William F. Ruddiman of Lamont-Doherty and John E. Kutzbach of the University of Wisconsin at Madison argue that the elevation of these landmasses has altered large-scale atmospheric circulation, producing the cooling trend that has marked the past 40 million years.

In any event, the dance of the plates has been going on for thousands of millions of years. Indeed, some workers have argued that Pangaea is only the most recent in a series of supercontinents that have formed and then fragmented once every several hundred million years. The evidence is tantalizingly scant: plate subduction serves as an excellent geologic shredder.

Contributing to the revolution in geology have been great improvements in seismology and—thanks to the digital computer—in the capacity to analyze the information that seismographs produce. One result is that earthquakes are much better understood. It may eventually be possible to predict them with an accuracy that would permit a rational, organized response. Another result is the ability to detect underground tests of nuclear weapons.

HAWAIIAN ISLANDS were made by the slow movement of a crustal plate over a hot spot. The photograph was taken during a space shuttle flight in 1986.

Nowhere is it more futile to draw a line between technology and science than in computing. Experiments in physics, suggested by the principles of quantum mechanics, yield beautifully precise knowledge of semiconducting materials. The work, coupled with technological developments, enables manufacturers to achieve a yearly improvement of 35 percent in the cost-performance ratio of solid state circuit components.

Advances in other sciences would have been difficult or impossible without the computer. Schrödinger, Einstein, Bohr, Thomson and Rutherford did not need computers to create theories or analyze experiments. But the teams of physicists who study the elementary particles that the Standard Model predicts would be lost without this extension of their intellectual capacity. Indeed, the computer has given science a third mode of inquiry, in addition to observation and experimentation: simulation.

Like scientific discoveries, the computer has transformed everyday life. It has become indispensable to the citizens of much of the industrial world who live in service economies. Without linked computers, the great global financial markets that every few hours handle a volume of money equal to a multiple of the U.S. national debt would be impossible; so would the air transportation system, or indeed the multinational corporation. The publishing industry has begun to use the machines in writing, editing and production. Some of the more enterprising players extend the electronic process to include the customers.

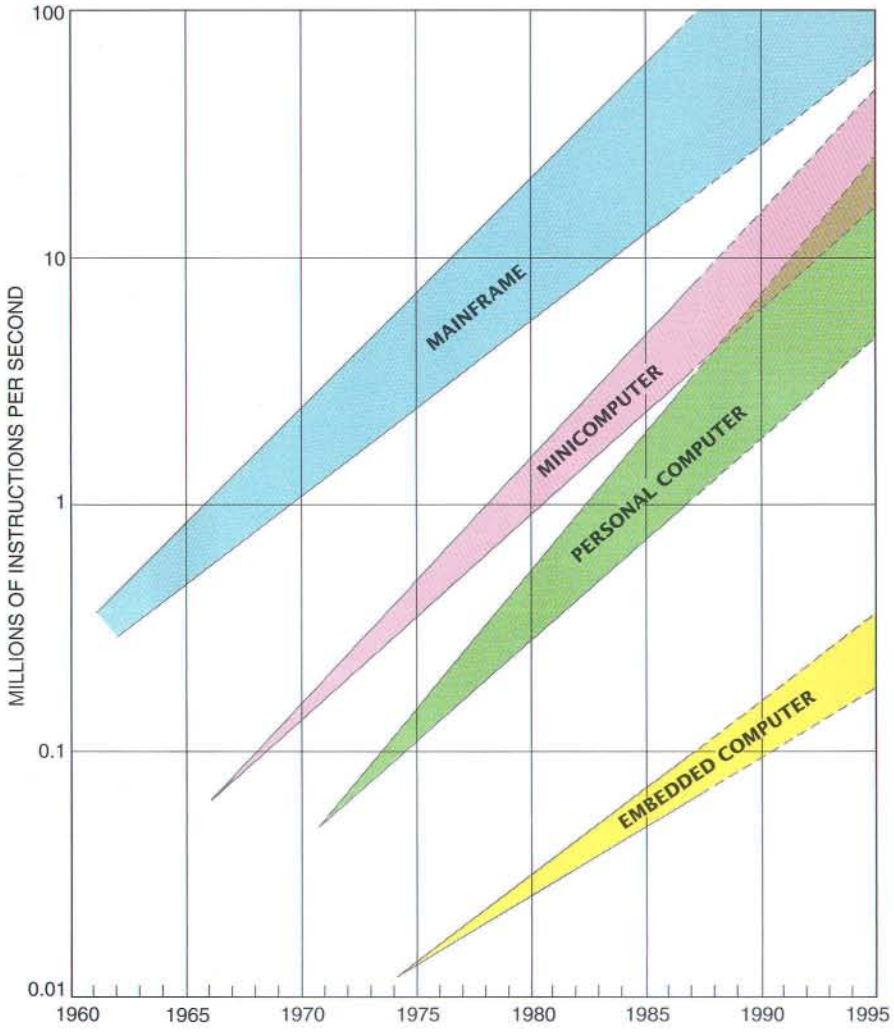
The computer is not a new idea. As early as 1812, Charles Babbage, an English inventor and mathematician, conceived of such a device and even succeeded in building part of it. In his autobiography, he speaks eloquently to his successors: "If, unwarned by my example, any man shall succeed in constructing an engine embodying in itself the whole of the executive department of mathematical analysis…I have no fear of leaving my reputation in his charge, for he alone will be able fully to appreciate the nature of my efforts and the value of their results."

The importance of linkage—the information and communications network—which first came to life in the form of the telegraph and then the telephone, inspired two other 19th-century figures. One was Karl Marx. In *Das Kapital* Marx observes that communications technologies such as the telegraph, the steamship and the steam railroad increase economic growth by accelerating flows of capital. His fellow visionary was the American novelist Nathaniel Hawthorne, who wrote in *The House of Seven Gables:* "Is it a fact…that by means of electricity, the world of matter has become a great nerve, vibrating thousands of miles in a breathless point of time? Rather, the round globe is a vast head, a brain, instinct with intelligence!" Hawthorne's crystal ball was not bad, considering that the year was 1851.

The invention of the transistor in 1948 at Bell Laboratories sprang the computer's rapid evolution. The transistor replaced the vacuum tube in electron control. It was miniaturized to the point where LSI (large-scale integration) and then very large scale integration (VLSI) of components on a small computing chip became possible. Today optical lithography can squeeze several million transistors onto a chip a square centimeter in area. Workers in the field now talk of GSI (gigascale integration), perhaps employing X-ray or electron-beam lithography to put a billion transistors on a chip.

Thus, the computer has shrunk as its capacity to do work has expanded. A mainframe machine in 1966 could perform as many as one million instructions per second. In 1975 that capacity was available from a minicomputer and in 1985 from a personal computer. Tiny embedded computers that control specific operations can be found in devices that range from copying machines to fly-by-wire aircraft.

Yet some problems (a detailed modeling of the development of the earth's climate, for example) thwart the most powerful conventional computer. Such challenges have given life to efforts to build supercomputers, machines capable of a trillion operations per second. The machines will embody parallel architecture. A parallel processing com-



GROWTH OF COMPUTING is charted for four types of machine. The colored bands define the range of computing power of each type of machine in each year.

puter executes the steps in a problem simultaneously rather than sequentially. At least one innovator has already begun to manufacture and sell a parallel processor that can perform billions of operations per second.

Meanwhile, back at the user interface, much work remains to be done. User naïveté—and frustration—has begat a whole technology of mice, icons, windows, electronic notepads, data gloves and virtual reality. Beyond this technology lies the promise of the truly transparent interface. Rapidly advancing speech production and recognition technology may be a major part of the answer.

Software is the soul of the computer. Writing it constitutes an intellectual challenge that has given rise to both an academic discipline and a robust, dynamic industry. The user of one of the 40 million personal computers in the U.S. can select from 20,000 software packages. And that is only the beginning. As David Gelernter of Yale University has put it, "Software systems are arising that...are information refineries that can transform mere facts into knowledge on a vast scale." Such systems may bring humankind the ultimate extension of our brains—artificial intelligence.

Such effects, profound as they are, can be strongly amplified. A network that links the machines enables users to create and share knowledge. Such a system must be able to carry traffic at the rate of a billion bits per second, far beyond the 64-kilobit capacity of the traditional telephone link. Enter photonics: the use of light rather than electric signals as the carrier of information. A photonic network carrying the electronic output of a computer as light can operate at a rate of a billion bits (gigabits) per second. Essential to such systems are solid state lasers and fiber optics—the thin glass "wires" that transmit the light signals.

Internet offers a glimpse of what life in such a cyberspace might be. Internet currently links 936 subsidiary networks, encompassing 175,000 computers in 35 countries. Like a vigorous kudzu vine, it has sprouted robustly and quite spontaneously from a seedling planted by the Department of Defense in 1969. A truly open and useful global system will require a higher degree of order. The job promises to be daunting.

There are also economic and policy questions to be considered: Who will pay for such a system? Who will own it? How, if at all, should it be regulated? Senator Al Gore of Tennessee has introduced a bill that would create a "national information highway" (his father as a senator was instrumental in establishing the national highway system). The administration reportedly might favor such a project.

When scientists talk to laypeople about science, they quite naturally tend to focus on what has been achieved rather than on the experience of working. Therefore, laypeople rarely see that doing science is a most human experience. Occasionally, though, we are offered a glimpse. Kary B. Mullis describes how he hit on the idea of the polymerase chain reaction as he drove north from San Francisco for the weekend, a friend asleep beside him. John V. Atanasoff, whom many historians (and at least one federal judge) credit with inventing the digital computer, said he experienced his conceptual breakthrough in an Illinois roadhouse where he had stopped for a drink, after driving 200 miles through the night from Ames, Iowa (the year was 1937).

Science is so human, in fact, that it is fun, even playful. At a symposium celebrating an eminent physicist's birthday, an astrophysicist inflated a balloon and appeared to pierce it with a conductor's baton without rupturing it, to demonstrate a topological point to his delighted colleagues. At luncheon an earnest journalist once asked John A. Wheeler, the great Princeton theoretician, what his favorite candidate for cold dark matter might be. "Baseballs," replied Wheeler, unhesitatingly. What is known about the structure and substance of the universe, Wheeler explained, so far outruns the reach of theory that wonder—like that felt by a child who wanders into a secret garden—rather than aggressive certainty, would provide the appropriate frame of mind. It is not a bad one to assume when looking at what science has achieved in this century or at the adventure that lies before us in the next one.

The telecommunications revolution of the 1990s

With photonics the driving force, multimedia telecommunications is moving ever closer to the era of Universal Information Services.

by John S. Mayo,
Senior Vice President, AT&T Bell Laboratories

Just as microelectronics propelled telecommunications for the last 20 years, photonics will drive the information revolution of the 1990s and beyond—and bring that revolution into the home.

The goal of the revolution is to have access to voice, data and images, in any combination, anywhere, at any time—and with convenience and economy. The goal is already set by the marketplace need for greatly enhanced information productivity, by the human desire for telepresence—a substitute for travel, and by our insatiable craving for entertainment.

These needs can be met by using information efficiently and effectively through such services as multi-media teleconferencing, distributed computing, remote interactive education programs, high-definition TV, and two-way switched video on demand.

The need for information productivity, the desire for telepresence, and the appetite for entertainment are powerful marketplace incentives for the evolution of an all-fiber network, including the last frontier—fiber to the home.

The cost of capacity

Entertainment represents a source of revenue that is needed to sustain and pay for the increased capacity of such a photonic network. It also represents a source for much of the information that the network would carry.

High-definition TV—HDTV—will require from 40 to 150 megabits-per-second data rates, and could be a major driving force for the extension of broadband capabilities to the home.

Photonics led to the progressive replacement of copper with fiber during the 1980s. During that decade, transmission made the transition to photonics.

During the 1990s, we will see the dynamic transition from electronics to photonics in other segments of telecommunications.

Photonics will probably never totally displace electronics, but will perform functions now thought to be beyond the capability of electronics. And it will begin to supplant electronics in many areas.

More specifically, during the 1990s, we will see the transition of switching to photonics. Arrays of lasers can beam their signals into optical media or even into free space to reach other photonic arrays. So photonics will enable connections without wires, freeing us from the constraints of physical wiring, and allowing us to reconfigure systems rapidly and at will.

Closing out the century

During the late 90s and beyond, we will see the transition of computing toward photonics. The wireless interconnections of photonics will provide nearly instantaneous computing. And because light beams do not interact with one

another, massively parallel computing architectures should be possible.

What is the status of photonics today? Photonics pervades virtually every aspect of transmission, from long-distance links to undersea cables to the local loops linking the customer with the central switching office. And optical fibers are replacing or complementing copper conductors under streets and oceans.

Photonic switching research now underway offers the promise of optical switching machines with terabit capacity—one trillion bits per second. This corresponds to almost 200 million facsimile terminals, to 660,000 video conferences, to nearly 20,000 studio-quality TV channels, or to 6,600 HDTV channels.

But photonics has already had a tremendous impact on the movement and management of information. Virtually all long-distance routes in the U.S. contain optical fiber systems operating at rates up to 1.7 gigabits per second, and that capacity is now being doubled to 3.4 gigabits—equal to nearly 50,000 simultaneous phone calls on a pair of fibers.

For transmission among business locations, we are seeing a network evolution toward fiber-based Local Area Networks (LANs) interconnecting computers and data bases.

There is also an evolution toward fiber-based Metropolitan Area Networks (MANs) to interconnect the various LANs.

Taking the last step

Fiber to the end user is becoming increasingly economical and will be the last step in fully achieving an all-fiber communications network. Such a network will have enormous bandwidth or information-carrying ability.

Moreover, photonic interconnections are being used inside electronic equipment and are being tested in the laboratory to connect individual silicon chips. And scientists at AT&T Bell Laboratories have fabricated a photonic digital processor, demonstrating the ability to replace electronics with photonics in elementary computing functions.

Laser transmission

In addition to low-loss fiber, viable semiconductor lasers were needed to launch photonics as the transmission technology of the 1980s. The first such lasers that operated at room temperature were fabricated in the 1970s. They were the size of grains of ordinary table salt, and could be easily coupled to optical fibers.

Today's most advanced semiconductor lasers are even smaller. Two million of them fit on a chip the size of a fingernail. These lasers emit light from their surfaces, instead of from their edges as do most lasers. After fabrication, that characteristic enables the surface-emitting lasers to be tested while they are still on a wafer. The surface-emitting property also allows them to be coupled easily in free-space architectures with other photonic components.

Eventually, a single chip will be able to emit millions of parallel beams of light. This light will be transmitted onto logic chips that, in turn, parallel-process the millions of beams and pass them onto other stages of logic.

Capacity versus distance

The overall capability of lightwave communications systems is measured by the product of two variables: the transmission data rate in megabits per second; and the distance in kilometers that signals can travel before they need regeneration.

That product or capability continues to double every year—driven by increasing data rates and decreasing fiber losses. This amazing pace will most likely continue for another two decades before we reach known physical limits. That means we can expect a thousand-fold improvement over the capabilities of today's most advanced lightwave systems.

Improvements in capability

Recent improvements in lightwave capability have come from using coherent technology and optical amplifiers.

Coherent systems offer the advantages of greater receiver sensitivity and longer fiber spans between regenerators. (They also offer greater receiver selectivity and ease of adding or dropping channels—in much the same way as FM radio stations are tuned in.)

The optical amplifier uses light to control light, and therefore eliminates the need for

high-speed electronics to regenerate signals. This amplifier is independent of bit rate. And just one amplifier will boost signals carried by many different wavelengths or "colors" of light.

This is especially important for systems using the capacity-building technique of wavelength division multiplexing—in which different "colors" of light are sent down the same fiber, each carrying different information.

Together, coherent technology and optical amplifiers will lead to fiber links across oceans—without the need for conventional regenerators.

Undersea cable systems

Undersea photonic cable systems are being installed in increasing numbers. Before the end of the decade, the equivalent of more than one million voice circuits will be placed under the oceans. About a dozen such major systems are either in service or planned.

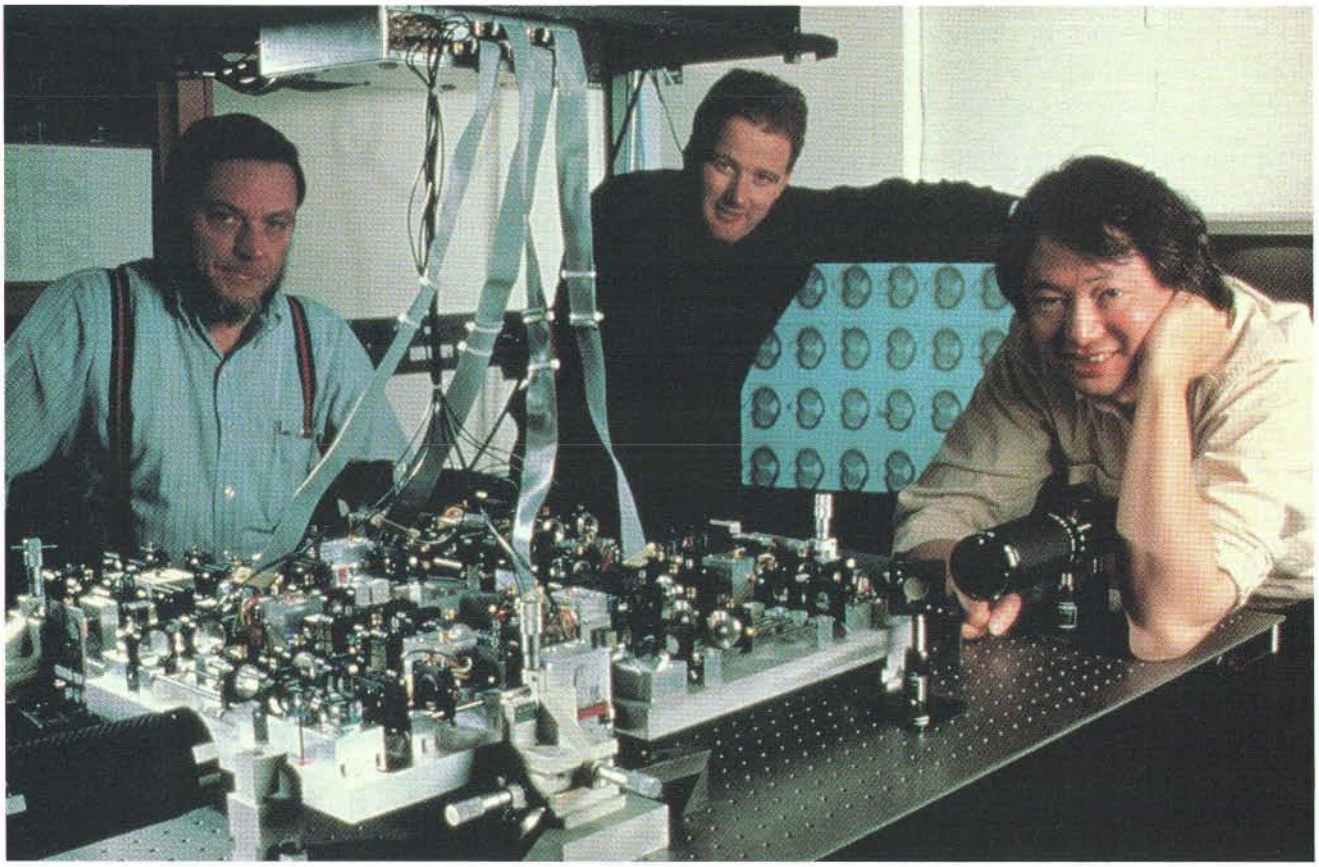
Fiber has clearly emerged as the medium of choice for transmission. To achieve the same information-carrying capacity as one fiber lightguide cable, we would need 155 reels of twisted-wire copper cable.

The lightguide cable is 23 times lighter than the copper cable. And its cross-sectional area is 36 times smaller. These two advantages—light weight and small size—make it easier to handle fiber cable in the field, and especially in crowded cable ducts. Fiber cable can also carry signals 28 miles before they need regeneration, compared to just over one mile for copper.

Moreover, the 28 miles of fiber cable need only about 100 two-way repeaters, while a copper system of equivalent information capacity would need about 20,000 two-way repeaters.

Taking fiber home

The next step—and the last step—toward an all-fiber network is to extend fiber to the home. This next move, extending fiber to the end user, has been stimulated by the desire to provide video services to end users, and by the introduction of the new digital format called Broadband Integrated Services Digital Network, or BISDN.



Members of the Optical Computing Research Department at AT&T Bell Laboratories in Holmdel, New Jersey, and the digital optical processor they helped build. Left to right: Bob LaMarche, Member of Technical Staff; Michael Prise, Member of Technical Staff; Alan Huang, head of the department.

Scott Hinton, head of the Optical Switching Department at AT&T Bell Laboratories, demonstrates the world's first photonic digital switching "fabric"—the part of the telecommunications system that switches voice, data or video from one place to another.

ISDN is an international network standard for integrating voice, data and image on the same line. And, in the most widely discussed form of BISDN, 622 megabits per second would be transmitted to each end user, who in turn could respond back to the network at 155 megabits per second. For this, the fiber link would terminate at the residence—and would provide BISDN capabilities.

Copper wire pairs cannot carry these higher bit-rate signals more than a few hundred feet. Only fiber has the appropriate transmission characteristics to carry very high bandwidth signals over extended distances—and to accommodate all video, data, telephony, and interactive data-base services.

Dealing with cost

Because the cost of installing fiber to the home is currently higher than with copper wiring, for some years fiber links will generally terminate at the curb, a few hundred feet from the residence, rather than at the residence itself. Sharing a fiber and its associated optoelectronic circuits among several residences will enable the costs of fiber access to be less than copper wire for new housing developments.

Once the fiber is deployed to the curb, the link from the curb to the residence can be made with either copper wire, copper coaxial cable, or fiber. The choice will depend on the kinds of services the customer selects. Higher-speed services can be made available to any customer by simply upgrading the link between the curb and the residence.

Growth of the residential fiber network is expected to be rapid. Within 15 years, the fiber network could grow to 100 million access lines. This rapid growth depends on the driving force of new video-based services, in addition to telephone service. Without such new services, fiber growth would be delayed by a full decade.

The anticipated use of the fiber network for video raises a tough issue. That use requires resolution of regulatory and ownership questions among local telephone companies, cable television providers, and other third-party service providers. Clearly, CATV could be a major driver of growth and full utilization of a fiber network.

The switch in switching

As we anticipate the use of the fiber network for video, it is also appropriate to consider the broad transition of switching to photonic technology.

During the 1960s, we saw the transition of switching to software control of electromechanical switch elements. In the 1970s, we saw the introduction of software control of electronic switch elements, in the form of all-digital toll switches. In the 1980s, we saw digital switching extended to local offices. And during the 1990s, we will see the transition to broadband switching—initially with software control of an electronic packet-switching fabric.

This stand-alone packet switch—for bursty types of data—will provide broadband ISDN, initially at data rates of 150 megabits per second. We will also see the transition to a broadband ISDN switching module for AT&T's 5ESS™ network switch, and perhaps for other switches as well. These AT&T switches will have software control of photonic switching fabrics.

The introduction of SEED

One of the most important emerging device structures for photonic switching and processing is the SEED or Self-Electro-optic Effect

Device. This device can control light with light, because its multilayer structure enables both detection and modulation of light.

As many as 2,000 symmetric SEEDs have been integrated into one chip. The switching speed of these early prototype devices is 2 megaHertz and is expected to increase by a factor of 100.

The symmetric SEED was crucial to the fabrication of the world's first photonic switch at Bell Labs. This uses light instead of electricity to switch information. The switch consists of an arrangement of lasers, lenses and S-SEED arrays on an optical bench. An array of 128 optical fibers feeds information in the form of light beams into the first of three S-SEED arrays. The 8-by-16 arrays are each about one-third the size of a grain of salt. That's a lot of connections in a very small area. The switch cascades the three arrays by connecting them in series one after the other. Any input port can be connected to a specific output port.

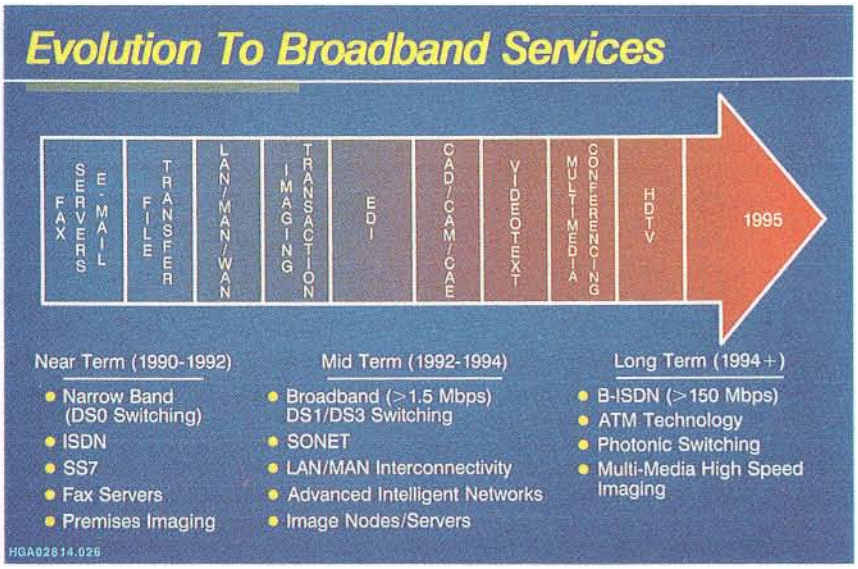


Chart above shows the near-term, mid-term and long-term evolution to broadband services.

The switch demonstrates this capability by unscrambling an input image to form an output display of lighted spots that form the letters "AT&T."

Such switching capacity would be well-matched to the growing capabilities of photonic transmission systems. Today, high-speed photonic information must be slowed down and converted to electronic format for processing in relatively slow electronic switching machines.

Digital photonic processing

In January of this year, Bell Labs scientists demonstrated the world's first digital photonic processor—which uses light to process information instead of electricity.

This experimental processor demonstrated the simple logic functions of counting and decoding. The processor consists of 4 arrays of 32 optical logic gates—formed from symmetric SEEDs. Each array occupies one corner of a square assemblage of lasers, lenses, and mirrors. The output of one array becomes the input for another array, and so on.

Information carried on the laser beams moves in and out of this four-stage cascade of logic gates, where it is processed.

The next stage in the evolution of photonic processors is to use VLSI technology to miniaturize and integrate optical components—to fabricate a compact system with fewer parts, requiring far less critical alignment of components.

The processor the Bell Labs scientists demonstrated is primitive.

But photonic processing offers the promise of computers with 1,000—even 10,000—times the processing power of their electronic counterparts.

Photonic transmission, photonic switching, and photonic computing.

These three technical forces will have a tremendous impact on the telecommunications revolution of the 1990s, moving us ever closer to the ultimate goal, Universal Information Services: The ability to provide voice, data and images, in any combination, anywhere, at any time, with convenience and economy.

John Mayo is the winner of the National Medal of Technology presented by President Bush.

What Is Matter?

The wave-particle dualism afflicting modern physics is best resolved in favor of waves, believes the author, but there is no clear picture of matter on which physicists can agree

by Erwin Schrödinger

Fifty years ago science seemed on the road to a clear-cut answer to the ancient question which is the title of this article. It looked as if matter would be reduced at last to its ultimate building blocks—to certain submicroscopic but nevertheless tangible and measurable particles. But it proved to be less simple than that. Today a physicist no longer can distinguish significantly between matter and something else. We no longer contrast matter with forces or fields of force as different entities; we know now that these concepts must be merged. It is true that we speak of "empty" space (that is, space free of matter), but space is never really empty, because even in the remotest voids of the universe there is always starlight—and *that* is matter. Besides, space is filled with gravitational fields, and according to Einstein gravity and inertia cannot very well be separated.

Thus, the subject of this article is in fact the total picture of space-time reality as envisaged by physics. We have to admit that our conception of material reality today is more wavering and uncertain than it has been for a long time. We know a great many interesting details, learn new ones every week. But to construct a clear, easily comprehensible picture on which all physicists would agree—that is simply impossible. Physics stands at a grave crisis of ideas. In the face of this crisis, many maintain that no objective picture of reality is possible. However, the optimists among us (of whom I consider myself one) look upon this view as a philosophical extravagance born of despair. We hope that the present fluctuations of thinking are only indications of an upheaval of old beliefs which in the end will lead to something better than the mess of formulas that today surrounds our subject.

Since the picture of matter that I am supposed to draw does not yet exist, since only fragments of it are visible, some parts of this narrative may be inconsistent with others. Like Cervantes's tale of Sancho Panza, who loses his donkey in one chapter but a few chapters later, thanks to the forgetfulness of the author, is riding the dear little animal again, our story has contradictions. We must start with the well-established concept that matter is composed of corpuscles or atoms, whose existence has been quite "tangibly" demonstrated by many beautiful experiments, and with Max Planck's discovery that energy also comes in indivisible units, called quanta, which are supposed to be transferred abruptly from one carrier to another.

But then Sancho Panza's donkey will return. For I shall have to ask you to believe neither in corpuscles as permanent individuals nor in the suddenness of the transfer of an energy quantum. Discreteness is present, but not in the traditional sense of discrete single particles, let alone in the sense of abrupt processes. Discreteness arises merely as a structure from the laws governing the phenomena. These laws are by no means fully understood; a probably correct analogue from the physics of palpable bodies is the way various partial tones of a bell derive from its shape and from the laws of elasticity to which, of themselves, nothing discontinuous adheres.

The idea that matter is made up of ultimate particles was advanced as early as the fifth century B.C. by Leucippus and Democritus, who called these particles atoms. The corpuscular theory of matter was lifted to physical reality in the theory of gases developed during the 19th century by James Clerk Maxwell and Ludwig Boltzmann. The concept of atoms and molecules in violent motion, colliding and rebounding again and again, led to full comprehension of all the properties of gases: their elastic and thermal properties, their viscosity, heat conductivity and diffusion. At the same time, it led to a firm foundation of the mechanical theory of heat, namely, that heat is the motion of these ultimate particles, which becomes increasingly violent with rising temperature.

Within one tremendously fertile decade at the turn of the century came the discoveries of X rays, of electrons, of the emission of streams of particles and other forms of energy from the atomic nucleus by radioactive decay, of the electric charges on the various particles. The masses of these particles, and of the atoms themselves, were later measured very precisely, and from

EDITOR'S NOTE

This article is condensed from a lecture entitled "Our Conception of Matter," delivered by Professor Schrödinger in 1952 at a conference in Geneva organized by Rencontres Internationales de Genève. The condensation is based on a translation by Sonja Bargmann, and it is published here with the kind permission of Editions de la Baconnière of Neuchâtel, Switzerland, who are publishing the full lecture in a volume called *L'homme devant la science,* presenting the proceedings of the conference.

ERWIN SCHRÖDINGER (1887–1961) was one of the founders of modern physics. For developing the theory of wave mechanics he shared the Nobel Prize in 1933 with the British physicist P. A. M. Dirac. Schrödinger, born in Vienna, came from the distinguished Austrian school of physics that produced Ernst Mach and Ludwig Boltzmann. He succeeded Max Planck in the chair of theoretical physics at the University of Berlin in 1927. Upon Hitler's rise to power he went to Dublin to join the Institute for Advanced Study, where he remained until 1956. In his later work he sought to combine the field theories of physics into a unified structure. He was also interested in more general unifications of science, and perhaps his most famous book was *What Is Life?*

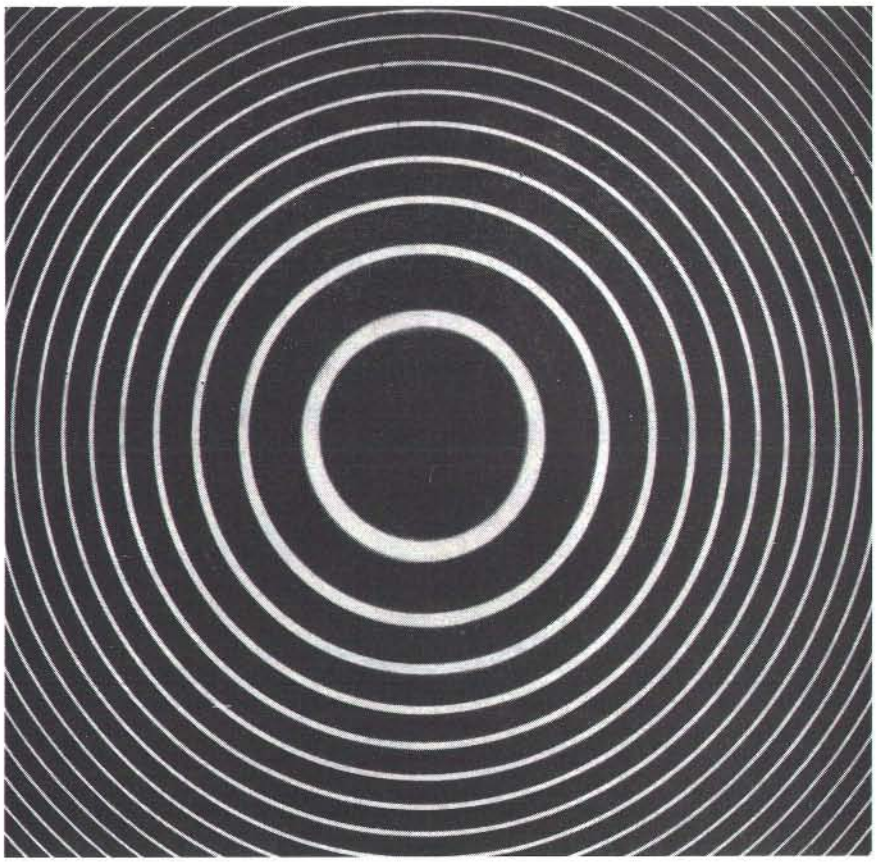
this was discovered the mass defect of the atomic nucleus as a whole. The mass of a nucleus is less than the sum of the masses of its component particles; the lost mass becomes the binding energy holding the nucleus firmly together. This is called the packing effect. The nuclear forces of course are not electrical forces—those are repellent—but are much stronger and act only within very short distances, about 10^{-13} centimeter.

Here I am already caught in a contradiction. Didn't I say at the beginning that we no longer assume the existence of force fields apart from matter? I could easily talk myself out of it by saying: "Well, the force field of a particle is simply considered a part of it." But that is not the fact. The established view today is rather that everything is at the same time both particle and field. Everything has the continuous structure with which we are familiar in fields, as well as the discrete structure with which we are equally familiar in particles. This concept is supported by innumerable experimental facts and is accepted in general, although opinions differ on details, as we shall see.

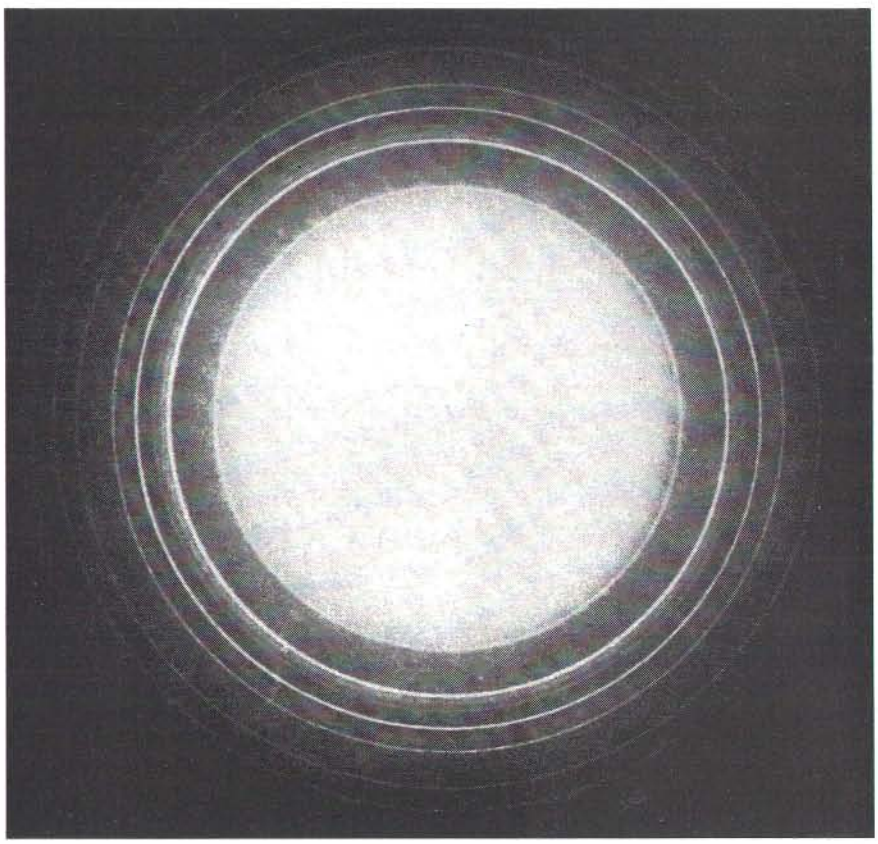
In the particular case of the field of nuclear forces, the particle structure is more or less known. Most likely, the continuous force field is represented by the so-called pi mesons. On the other hand, the protons and neutrons, which we think of as discrete particles, also have a continuous wave structure, as is shown by the interference patterns they form when diffracted by a crystal. The difficulty of combining these two so very different character traits in one mental picture is the main stumbling block that causes our conception of matter to be so uncertain.

Neither the particle concept nor the wave concept is hypothetical. The tracks in a photographic emulsion or in a Wilson cloud chamber leave no doubt of the behavior of particles as discrete units. The artificial production of nuclear particles is being attempted right now with terrific expenditure, defrayed in the main by the various state ministries of defense. It is true that one cannot kill anybody with one such racing particle, or else we should all be dead by now. But their study promises, indirectly, a hastened realization of the plan for the annihilation of mankind which is so close to all our hearts.

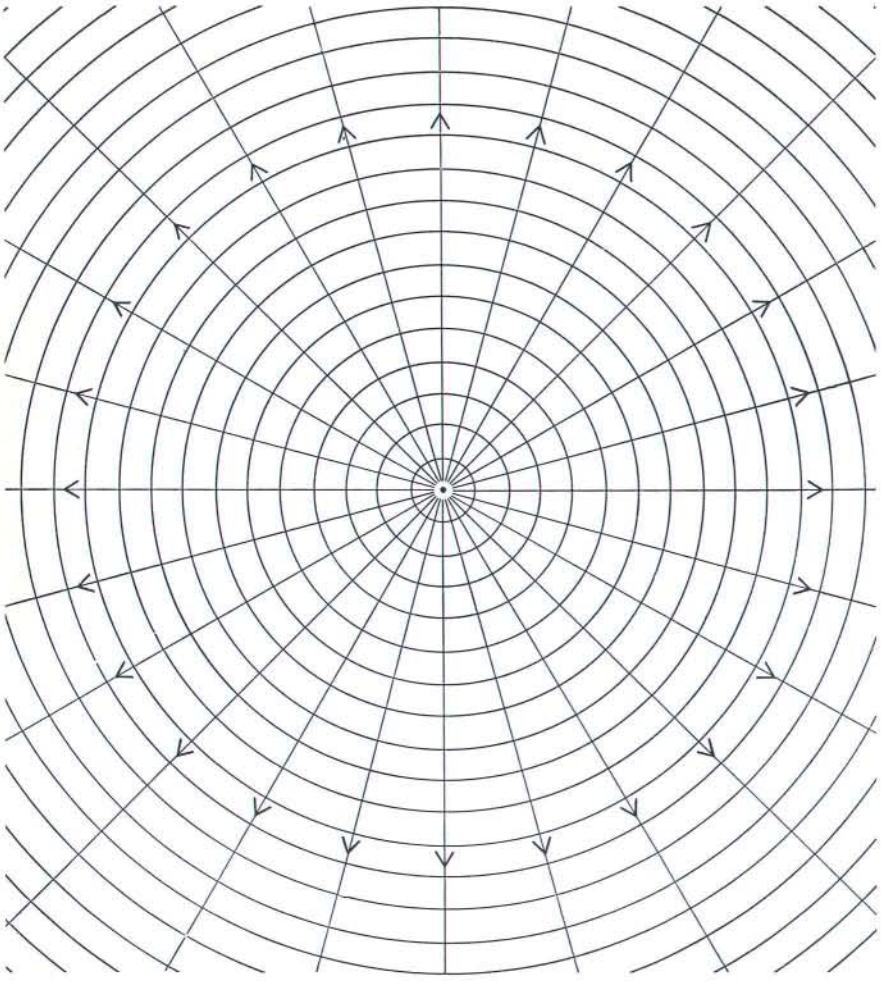
You can easily observe particles yourself by looking at a luminous numeral of your wrist watch in the dark with a magnifying glass. The luminosity surges and undulates, just as a lake



LIGHT INTERFERENCE pattern, showing the wave nature of light, was produced at the National Bureau of Standards, using light from mercury vapor.



ELECTRON INTERFERENCE pattern from a crystal diffraction experiment at the Radio Corporation of America Laboratories shows that electrons are waves.



WAVE DIAGRAM in two dimensions shows wave fronts (*circles*) and wave "normals" or "rays" (*arrows*). Three-dimensional fronts would resemble layers in an onion.

sometimes twinkles in the sun. The light consists of sparklets, each produced by a so-called alpha particle (helium nucleus) expelled by a radioactive atom which in this process is transformed into a different atom. A specific device for detecting and recording single particles is the Geiger-Müller counter. In this short résumé I cannot possibly exhaust the many ways in which we can observe single particles.

Now to the continuous field or wave character of matter. Wave structure is studied mainly by means of diffraction and interference—phenomena that occur when wave trains cross each other. For the analysis and measurement of light waves the principal device is the ruled grating, which consists of a great many fine, parallel, equidistant lines, closely engraved on a specular metallic surface. Light impinging from one direction is scattered by them and collected in different directions depending on its wavelength. But even the finest ruled gratings we can produce are too coarse to scatter the very much shorter waves associated with matter. The fine lattices of crystals, however, which Max von Laue first used as gratings to analyze the very short X rays, will do the same for "matter waves." Directed at the surface of a crystal, high-velocity streams of particles manifest their wave nature. With crystal gratings, physicists have diffracted and measured the wavelengths of electrons, neutrons and protons.

What does Planck's quantum theory have to do with all this? Planck told us in 1900 that he could comprehend the radiation from red-hot iron, or from an incandescent star such as the sun, only if this radiation was produced in discrete portions and transferred in such discrete quantities from one carrier to another (for example, from atom to atom). This was extremely startling, because up to that time energy had been a highly abstract concept. Five years later Einstein told us that energy has mass and mass is energy; in other words, that they are one and the same. Now the scales begin to fall from our eyes: our dear old atoms, corpuscles, particles are Planck's energy quanta. *The carriers of those quanta are themselves quanta.* One gets dizzy. Something quite fundamental must lie at the bottom of this, but it is not surprising that the secret is not yet understood. After all, the scales did not fall suddenly. It took 20 or 30 years. And perhaps they still have not fallen completely.

The next step was not quite so far-reaching, but important enough. By an ingenious and appropriate generalization of Planck's hypothesis, Niels Bohr taught us to understand the line spectra of atoms and molecules and how atoms were composed of heavy, positively charged nuclei with light, negatively charged electrons revolving around them. Each small system—atom or molecule—can harbor only definite discrete energy quantities, corresponding to its nature. In transition from a higher to a lower "energy level," it emits the excess energy as a radiation quantum of definite wavelength, inversely proportional to the quantum given off. This means that a quantum of given magnitude manifests itself in a periodic process of definite frequency that is directly proportional to the quantum; the frequency equals the energy quantum divided by the famous Planck's constant, h.

According to Einstein, a particle has the energy mc^2, m being the mass of the particle and c the velocity of light. In 1925 Louis de Broglie drew the inference, which rather suggests itself, that a particle might have associated with it a wave process of frequency mc^2 divided by h. The particle for which he postulated such a wave was the electron. Within two years the "electron waves" required by his theory were demonstrated by the famous electron diffraction experiment of C. J. Davisson and L. H. Germer. This was the starting point for the cognition that everything—anything at all—is simultaneously particle and wave field. Thus, de Broglie's dissertation initiated our uncertainty about the nature of matter. Both the particle picture and the wave picture have truth value, and we cannot give up either one or the other. But we do not know how to combine them.

That the two pictures are connected is known in full generality with great precision and down to amazing details. But concerning the unification to a single, concrete, palpable picture, opinions are so strongly divided that a great many deem it alto-

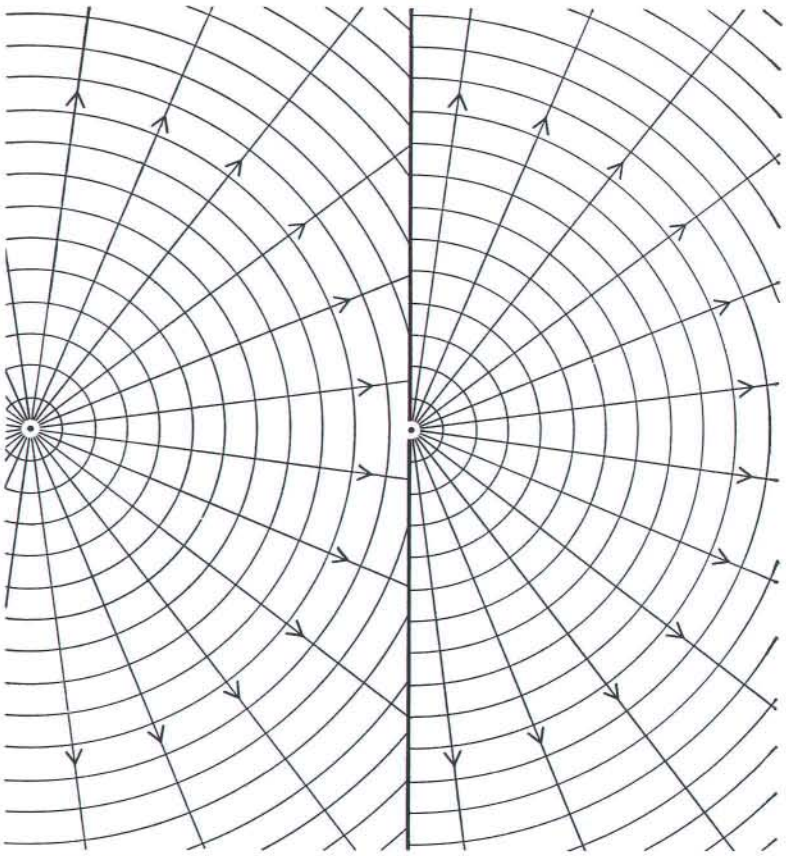
gether impossible. I shall briefly sketch the connection. But do not expect that a uniform, concrete picture will emerge before you, and do not blame the lack of success either on my ineptness in exposition or your own denseness—nobody has yet succeeded.

One distinguishes two things in a wave. First, a wave has a front, and a succession of wave fronts forms a system of surfaces like the layers of an onion. A two-dimensional analogue is the beautiful wave circles that form on the smooth surface of a pond when a stone is thrown in. The second characteristic of a wave, less intuitive, is the path along which it travels—a system of imagined lines perpendicular to the wave fronts. These lines are known as the wave "normals" or "rays."

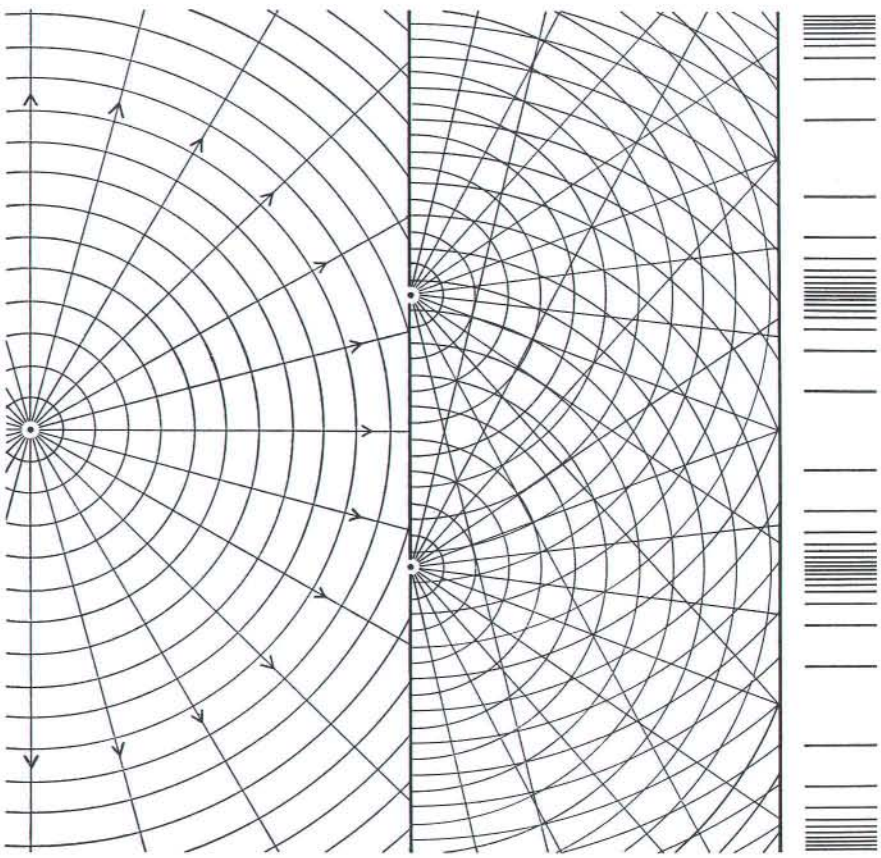
We can make the provisional assertion that these rays correspond to the trajectories of particles. Indeed, if you cut a small piece out of a wave, approximately 10 or 20 wavelengths along the direction of propagation and about as much across, such a "wave packet" would actually move along a ray with exactly the same velocity and change of velocity as we might expect from a particle of this particular kind at this particular place, taking into account any force fields acting on the particle.

Here I falter. For what I must say now, though correct, almost contradicts this provisional assertion. Although the behavior of the wave packet gives us a more or less intuitive picture of a particle, which can be worked out in detail (for example, the momentum of a particle increases as the wavelength decreases; the two are inversely proportional), yet for many reasons we cannot take this intuitive picture quite seriously. For one thing, it is, after all, somewhat vague, the more so the greater the wavelength. For another, quite often we are dealing not with a small packet but with an extended wave. For still another, we must also deal with the important special case of very small "packelets" which form a kind of "standing wave" that can have no wave fronts or wave normals.

One interpretation of wave phenomena extensively supported by experiments is this: at each position of a uniformly propagating wave train, there is a twofold structural connection of interactions, which may be distinguished as "longitudinal" and "transversal." The transversal structure is that of the wave fronts and manifests itself in diffraction and interference experiments; the longitudinal structure is that of the wave normals and manifests itself in the observation of single particles. However, these concepts of longitudi-



DIFFRACTION is characteristic of waves. When a wave (*left*) comes to a barrier with a small hole, it diffracts around the edges, thereby forming a new wave (*right*).



INTERFERENCE is also evidence of waves. Its characteristic pattern is formed when rays interact. For light waves, it is bright and dark bands on a screen (*right*).

nal and transversal structures are not sharply defined and absolute, since the concepts of wave front and wave normal are not, either.

The interpretation breaks down completely in the special case of the standing waves mentioned above. Here the whole wave phenomenon is reduced to a small region of the dimensions of a single or very few wavelengths. You can produce standing water waves of a similar nature in a small basin if you dabble with your finger rather uniformly in its center, or else just give it a little push so that the water surface undulates. In this situation we are not dealing with uniform wave propagation; what catches the interest are the normal frequencies of these standing waves. The water waves in the basin are an analogue of a wave phenomenon associated with electrons, which occurs in a region just about the size of the atom. The normal frequencies of the wave group washing around the atomic nucleus are universally found to be exactly equal to Bohr's atomic "energy levels" divided by Planck's constant *h*. Thus, the ingenious yet somewhat artificial assumptions of Bohr's model of the atom, as well as of the older quantum theory in general, are superseded by the far more natural idea of de Broglie's wave phenomenon. The wave phenomenon forms the "body" proper of the atom. It takes the place of the individual pointlike electrons, which in Bohr's model are supposed to swarm around the nucleus. Such pointlike single particles are completely out of the question within the atom, and if one still thinks of the nucleus itself in this way, one does so quite consciously for reasons of expediency.

What seems to me particularly important about the discovery that "energy levels" are virtually nothing but the frequencies of normal modes of vibration is that as a result one can do without the assumption of sudden transitions, or quantum jumps, since two or more normal modes may very well be excited simultaneously. The discreteness of the normal frequencies fully suffices—so I believe—to support the considerations from which Planck started and many similar and just as important ones—I mean, in short, to support all of quantum thermodynamics.

The theory of quantum jumps is becoming more and more unacceptable, at least to me personally, as the years go on. Its abandonment has, however, far-reaching consequences. It means that one must give up entirely the idea of the exchange of energy in well-defined quanta and replace it with the concept of resonance between vibrational frequencies. Yet we have seen that because of the identity of mass and energy, we must consider the particles themselves as Planck's energy quanta. This is at first frightening. For the substituted theory implies that we can no longer consider the individual particle as a well-defined permanent entity.

That it is, in fact, no such thing can be reasoned in other ways. For one thing, there is Werner Heisenberg's famous uncertainty principle, according to which a particle cannot simultaneously have a well-defined position and a sharply defined velocity. This uncertainty implies that we cannot be sure that the same particle could ever be observed twice. Another conclusive reason for not attributing identifiable sameness to individual particles is that we must obliterate their individualities whenever we consider two or more interacting particles of the same kind, for example, the two electrons of a helium atom. Two situations that are distinguished only by the interchange of the two electrons must be counted as one and the same; if they are counted as *two* equal situations, nonsense obtains. This circumstance holds for any kind of particle in arbitrary numbers without exception.

Most theoreticians will probably accept the foregoing reasoning and admit that the individual particle is not a well-defined permanent entity of detectable identity or sameness. Nevertheless, this inadmissible concept of the individual particle continues to play a large role in their ideas and discussions. Even deeper rooted is the belief in "quantum jumps," which is now surrounded with a highly abstruse terminology whose commonsense meaning is often difficult to grasp. For instance, an important word in the standing vocabulary of quantum theory is "probability," referring to transition from one level to another. But, after all, one can speak of the probability of an event only assuming that, occasionally, it actually occurs. If it does occur, the transition must be sudden, since intermediate stages are disclaimed. Moreover, if it takes time, it might be interrupted halfway by an unforeseen disturbance. This possibility leaves one completely at sea.

The wave versus corpuscle dilemma is supposed to be resolved by asserting that the wave field merely serves for the computation of the probability of finding a particle of given properties at a given position if one looks for it there. But once one deprives the waves of reality and assigns them only a kind of informative role, it becomes very difficult to understand the phenomena of interference and diffraction on the basis of the combined action of discrete single particles. It seems easier to explain particle tracks in terms of waves than to explain the wave phenomenon in terms of corpuscles.

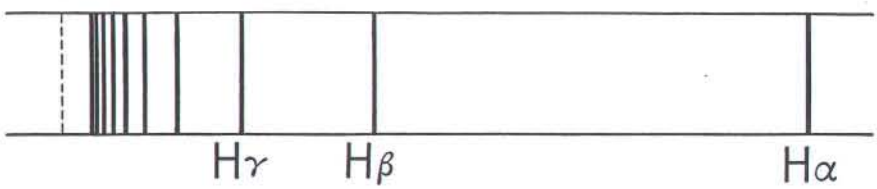
"Real existence" is, to be sure, an expression that has been virtually chased to death by many philosophical hounds. Its simple, naive meaning has almost become lost to us. Therefore, I want to recall something else. I spoke of a corpuscle's not being an individual. Properly speaking, one never observes the same particle a second time—very much as Heraclitus says of the river. You cannot mark an electron, you cannot paint it red. Indeed, you must not even *think* of it as marked; if you do, your "counting" will be false and you will get wrong results at every step—for the structure of line spectra, in thermodynamics and elsewhere. A wave, on the other hand, can easily be imprinted with an individual structure by which it can be recognized beyond doubt. Think of the beacon fires that guide ships at sea. The light shines according to a definite code; for example, three seconds light, five seconds dark, one second light, another pause of five seconds, and again light for three seconds—the skipper knows that is San Sebastian. Or you talk by wireless telephone with a friend across the Atlantic; as soon as he says, "Hello there, Edward Meier speaking," you know that his voice has imprinted on the radio wave a structure which can be distinguished from any other. But one does not have to go that far. If your wife calls, "Francis!" from the garden, it is exactly the same thing, except that the structure is printed on sound waves and the trip is shorter (though it takes somewhat longer than the journey of radio waves across the Atlantic). All our verbal communication is based on imprinted individual wave structures. And, according to the same principle, what a wealth of details is transmitted to us in rapid succession by the movie or the television picture!

This characteristic, the individuality of the wave phenomenon, has already been found to a remarkable extent in the very much finer waves of particles. One example must suffice. A limited volume of gas, say, helium, can be thought of either as a collection of many helium atoms or as a superposition of elementary wave trains of matter waves. Both views lead to the

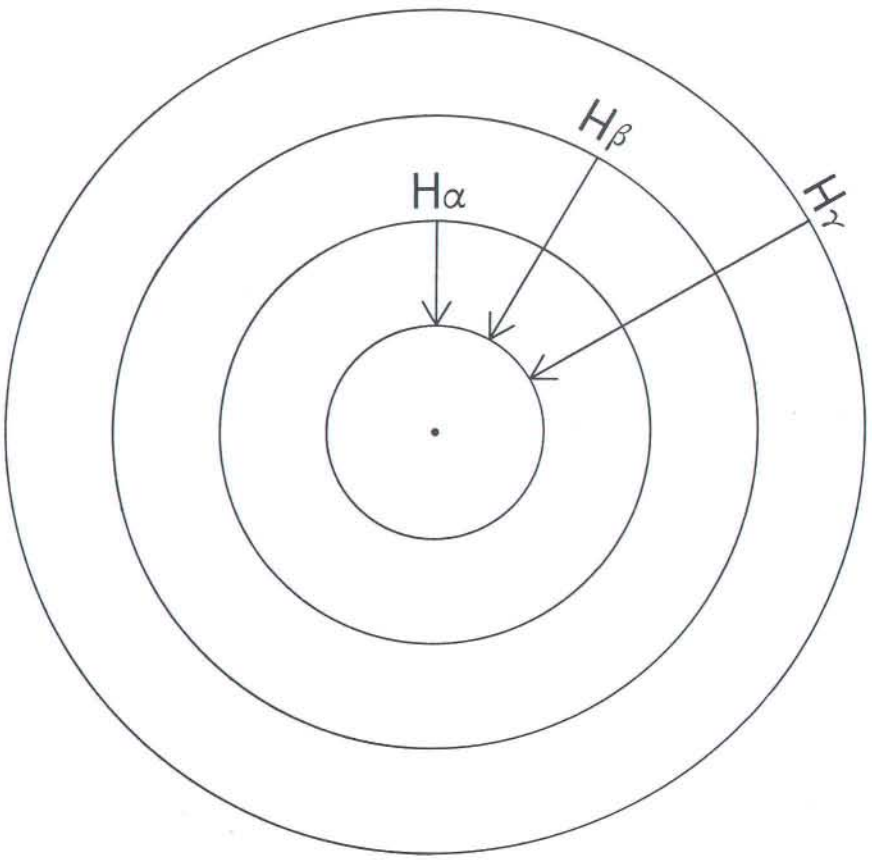
same theoretical results as to the behavior of the gas upon heating, compression and so on. But when you attempt to apply certain somewhat involved enumerations to the gas, you must carry them out in different ways according to the mental picture with which you approach it. If you treat the gas as consisting of particles, no individuality must be ascribed to them. If, however, you concentrate on the matter wave trains instead of on the particles, every one of the wave trains has a well-defined structure that is different from that of any other. It is true that there are many pairs of waves so similar to each other that they could change roles without any noticeable effect on the gas. But if you should count the very many similar states formed in this way as merely a single one, the result would be quite wrong.

In spite of everything, we cannot completely banish the concepts of quantum jump and individual corpuscle from the vocabulary of physics. We still require them to describe many details of the structure of matter. How can one ever determine the weight of a carbon nucleus and of a hydrogen nucleus, each to the precision of several decimals, and detect that the former is somewhat lighter than the 12 hydrogen nuclei combined in it, without accepting for the time being the view that these particles are something quite concrete and real? This view is so much more convenient than the roundabout consideration of wave trains that we cannot do without it, just as the chemist does not discard his valence-bond formulas, although he fully realizes that they represent a drastic simplification of a rather involved wave-mechanical situation.

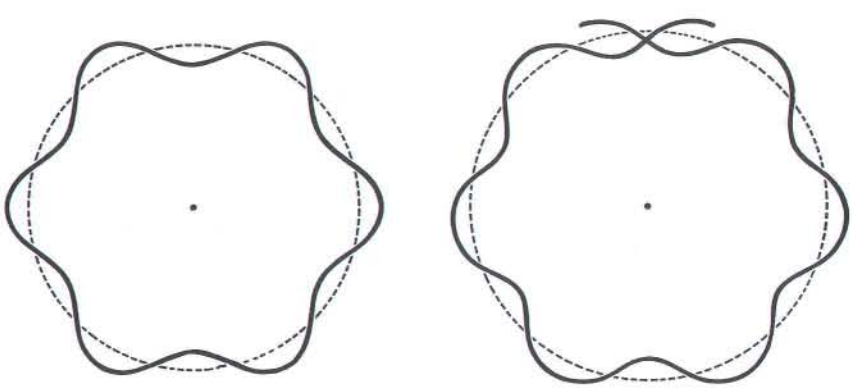
If you finally ask me: "Well, what *are* these corpuscles, really?" I ought to confess honestly that I am almost as little prepared to answer that as to tell where Sancho Panza's second donkey came from. At the most, it may be permissible to say that one can think of particles as more or less temporary entities within the wave field whose form and general behavior are nevertheless so clearly and sharply determined by the laws of waves that many processes take place *as if* these temporary entities were substantial permanent beings. The mass and the charge of particles, defined with such precision, must then be counted among the structural elements determined by the wave laws. The conservation of charge and mass in the large must be considered as a statistical effect, based on the "law of large numbers."



HYDROGEN SPECTRUM expresses the behavior of a fundamental constituent of matter, the electron. Shown above is a part of the Balmer series of spectral lines. Each line is the result of a change in energy of the atom's electron.



BOHR THEORY explained spectral lines of hydrogen by postulating a pointlike electron revolving around the nucleus in an orbit. In falling from one to another, the electron emits light energy whose wavelength is that of one of the spectral lines.



WAVE MECHANICS sees the electron not as a point mass but as a standing wave washing to and fro in the atom. Some modes of vibration are possible (*left*), others are not (*right*). The possible modes match the Bohr theory's possible energy levels.

Unified Theories of Elementary Particle Interaction

Physicists now invoke four distinct kinds of interaction, or force, to describe physical phenomena. According to a new theory, two, and perhaps three, of the forces are seen to have an underlying identity

by Steven Weinberg

One of humankind's enduring hopes has been to find a few simple general laws that would explain why nature, with all its seeming complexity and variety, is the way it is. At the present moment, the closest we can come to a unified view of nature is a description in terms of elementary particles and their mutual interactions. All ordinary matter is composed of just those elementary particles that happen to possess both mass and (relative) stability: the electron, the proton and the neutron. To these must be added the particles of zero mass: the photon, or quantum of electromagnetic radiation; the neutrino, which plays an essential role in certain kinds of radioactivity; and the graviton, or quantum of gravitational radiation. (The graviton interacts too weakly with matter for it to have been observed yet, but there is no serious reason to doubt its existence.) A few additional short-lived particles can be found in cosmic rays, and with particle accelerators it is possible to create a vast number of even shorter-lived species [*see top illustration on page 24*].

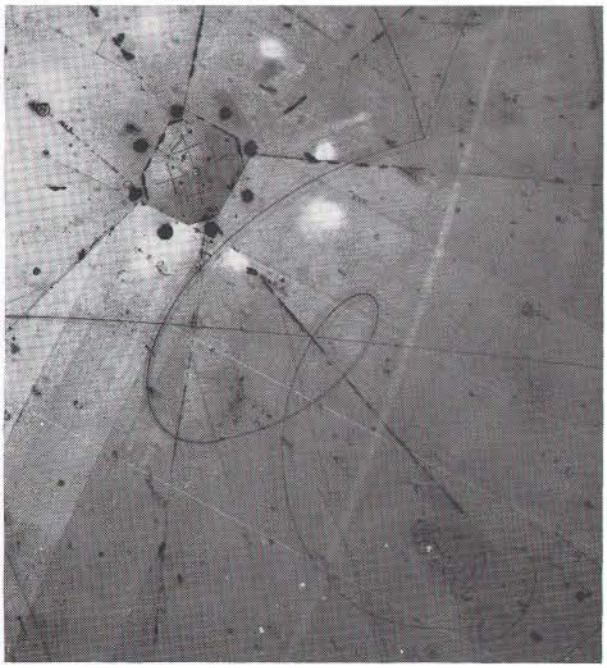
STEVEN WEINBERG is Josey Professor of Science at the University of Texas at Austin and senior consultant at the Smithsonian Astrophysical Laboratoy. Before joining the Texas faculty, he had served at the Massachusetts Institute of Technology, the University of California, Berkeley, Columbia University and Harvard University, where he was Higgins Professor of Physics. Weinberg shared the Nobel Prize in Physics in 1979 with Sheldon L. Glashow and Abdus Salam; the prize was awarded for their complementary work in formulating a theory that encompasses both the electromagnetic and the weak interaction of elementary particles. Weinberg's extracurricular interests include medieval history and arms control; he has served as a consultant to the U.S. Arms Control and Disarmament Agency. He is the author of, among other books, *The First Three Minutes: A Modern View of the Origin of the Universe* and *The Discovery of Subatomic Particles.*

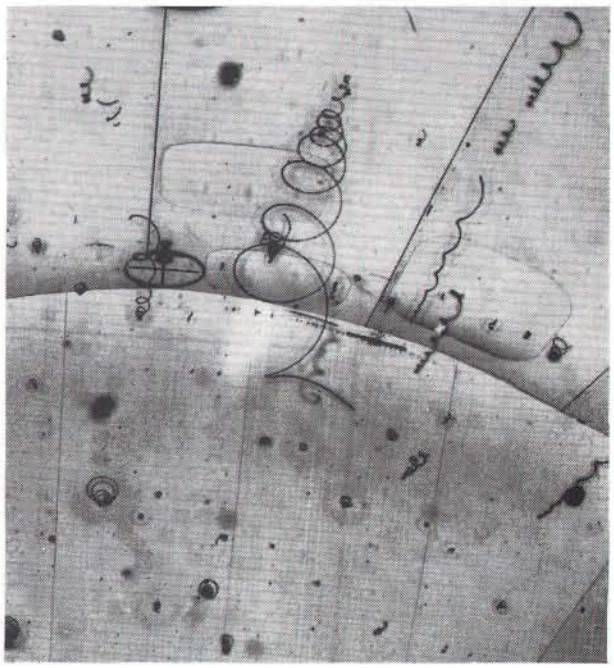
Although the various particles differ widely in mass, charge, lifetime as well as in other ways, they all share two attributes that qualify them as being "elementary." First, as far as we know, any two particles of the same species are, except for their position and state of motion, absolutely identical, whether they occupy the same atom or lie at opposite ends of the universe. Second, there is not now any successful theory that explains the elementary particles in terms of more elementary constituents, in the sense that the atomic nucleus is understood to be composed of protons and neutrons and the atom is understood to be composed of a nucleus and electrons. It is true that the elementary particles behave in some respects as if they were composed of still more elementary constituents, which are named quarks, but in spite of strenuous efforts it has so far proved to be impossible to break particles into quarks.

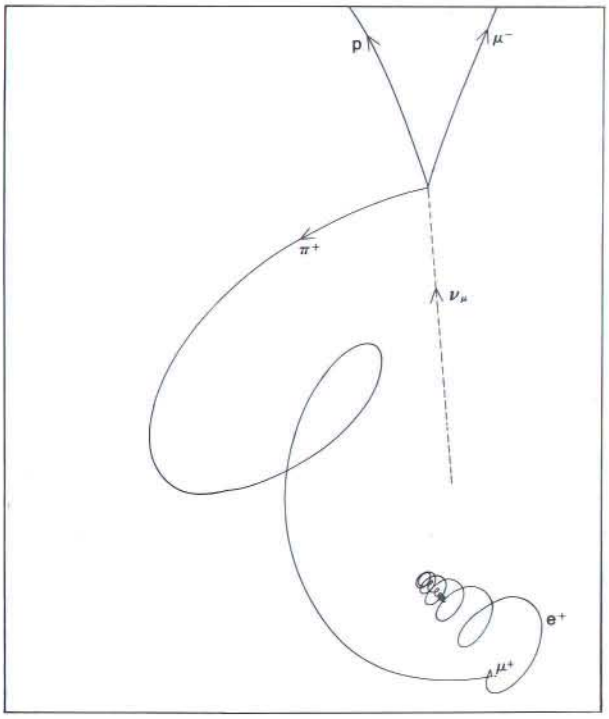
For all the bewildering variety of the elementary particles, their interactions with one another appear to be confined to four broad categories [*see bottom illustration on page 24*]. The most familiar are gravitation and electromagnetism, which, because of their long range, are experienced in the everyday world. Gravity holds our feet on the ground and the planets in their orbits. Electromagnetic interactions of electrons and atomic nuclei are responsible for all the familiar chemical and physical properties of ordinary solids, liquids and gases. Next, both in range and familiarity, are the "strong" interactions, which hold protons and neutrons together in the atomic nucleus. The strong forces are limited in range to about 10^{-13} centimeter and so are quite insignificant in ordinary life, or even on the scale (10^{-8} centimeter) of the atom. Least familiar are the "weak" interactions. They are of such short range (less than 10^{-15} centimeter) and are so weak that they do not seem to play a role in holding anything together. Rather they are manifested only in certain kinds of collisions or decay processes that, for whatever reason, cannot be mediated by the strong, electromagnetic or gravitational interactions. The weak interactions are not, however, irrelevant to human affairs. They provide the first step in the chain of thermonuclear reactions in the sun, a step in which two protons fuse to form a deuterium nucleus, a positron and a neutrino.

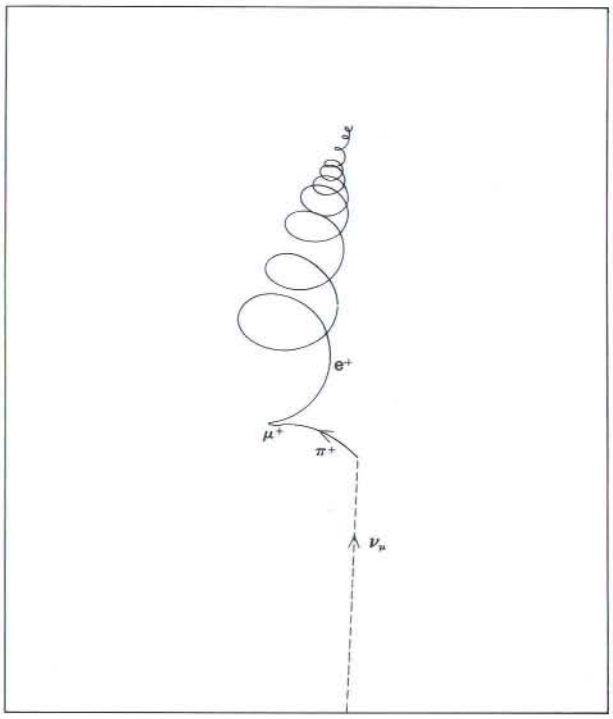
From this brief outline, one can see that a certain measure of unification has been achieved in making sense of the world. We are still faced, however, with the enormous problem of accounting for the baffling variety of elementary particle types and interactions. Our prospects for further progress would be truly discouraging were it not for the guidance we receive from two great products of 20th-century physics: the development of quantum field theory and the recognition of the fundamental role of symmetry principles.

Reprinted from the July 1974 issue









EVIDENCE FOR NEUTRAL CURRENTS, the existence of which would support theories showing a connection between electromagnetic interactions and weak interactions, was recently obtained in an experiment conducted at the Argonne National Laboratory with a neutrino beam from the zero-gradient synchrotron and with a 12-foot bubble chamber filled with liquid hydrogen. The bubble-chamber photograph at the left and the map below it show an example of a familiar kind of charged-current process ($\nu_\mu + p \rightarrow \mu^- + p + \pi^+$), in which a unit of electric charge is exchanged between leptons (ν_μ, μ^-) and other particles. The photograph at the right and the map below it show an example of a neutral-current process ($\nu_\mu + p \rightarrow \nu_\mu + n + \pi^+$) distinguished by the absence of outgoing negative muon (μ^-) or proton (p) tracks. In such photographs, tracks are left only by charged particles, so that the incoming neutrino (ν_μ) and the outgoing neutrino and neutron (n) in the neutral-current process are invisible. Moreover, the bubble chamber is subjected to an intense magnetic field, which causes charged particles to follow curved tracks, clockwise for negative charge and counterclockwise for positive charge. In both of these photographs the positive pion (π^+) is seen to decay into a positive muon (μ^+), which then decays into a positive electron (e^+), visible as a tightly wound spiral. This experiment is more recent than similar ones that have been conducted at the European laboratory for particle physics (CERN) and at the Fermi National Accelerator Laboratory. It provides the first evidence for the specific neutral-current reactions $\nu_\mu + p \rightarrow \nu_\mu + n + \pi^+$ and $\nu_\mu + p \rightarrow \nu_\mu + p + \pi^0$.

		PARTICLE	SYMBOL	CHARGE	MASS (10^6 ELECTRON VOLTS)	LIFETIME (SECONDS)
		PHOTON	γ	0	0	∞
LEPTONS		NEUTRINO	$\nu_e\ \overline{\nu}_e$	0	0	∞
			$\nu_\mu\ \overline{\nu}_\mu$	0	0	∞
		ELECTRON	$e^\pm$	$\pm e$	0.511	∞
		MUON	$\mu^\pm$	$\pm e$	105.66	2.199×10^{-6}
HADRONS	MESONS	PION	$\pi^\pm$	$\pm e$	139.57	2.602×10^{-8}
			π°	0	134.97	0.84×10^{-16}
		KAON	$K^\pm$	$\pm e$	493.71	1.237×10^{-8}
			K°	0	497.71	0.882×10^{-10}
		ETA	η	0	548.8	2.50×10^{-17}
	BARYONS	PROTON	$p\ \overline{p}$	$\pm e$	938.259	∞
		NEUTRON	$n\ \overline{n}$	0	939.553	918
		LAMBDA HYPERON	$\Lambda\ \overline{\Lambda}$	0	1,115.59	2.521×10^{-10}
		SIGMA HYPERON	$\Sigma^+\ \overline{\Sigma}^+$	$\pm e$	1,189.42	8.00×10^{-11}
			$\Sigma^\circ\ \overline{\Sigma}^\circ$	0	1,192.48	$<10^{-14}$
			$\Sigma^-\ \overline{\Sigma}^-$	$\pm e$	1,197.34	1.484×10^{-10}
		CASCADE HYPERON	$\Xi^\circ\ \overline{\Xi}^\circ$	0	1,314.7	2.98×10^{-10}
			$\Xi^-\ \overline{\Xi}^-$	$\pm e$	1,321.3	1.672×10^{-10}
		OMEGA HYPERON	$\Omega^-\ \overline{\Omega}^-$	$\pm e$	1,672	1.3×10^{-10}

PARTIAL LIST OF OBSERVED ELEMENTARY PARTICLES identifies all those with lifetimes greater than 10^{-20} second. Apart from the photon, all the observed particles fall into two broad families: leptons and hadrons. The leptons are either massless or low-mass particles that do not take part in the "strong" interactions; the hadrons are heavier and do take part. Hadrons are further divided into mesons and baryons according to rotational angular momentum and other properties. A symbol with a bar above it denotes an antiparticle. Neutrinos and antineutrinos are of two types: electron-type, ν_e, and muon-type, ν_μ. In three cases (photon, neutral pion and eta meson), the particle is its own antiparticle. Charges are given in units of charge, *e*, on the electron, equal to 1.602×10^{-19} coulomb. Masses are in energy units; one million electron volts (MeV) equals 1.783×10^{-27} gram.

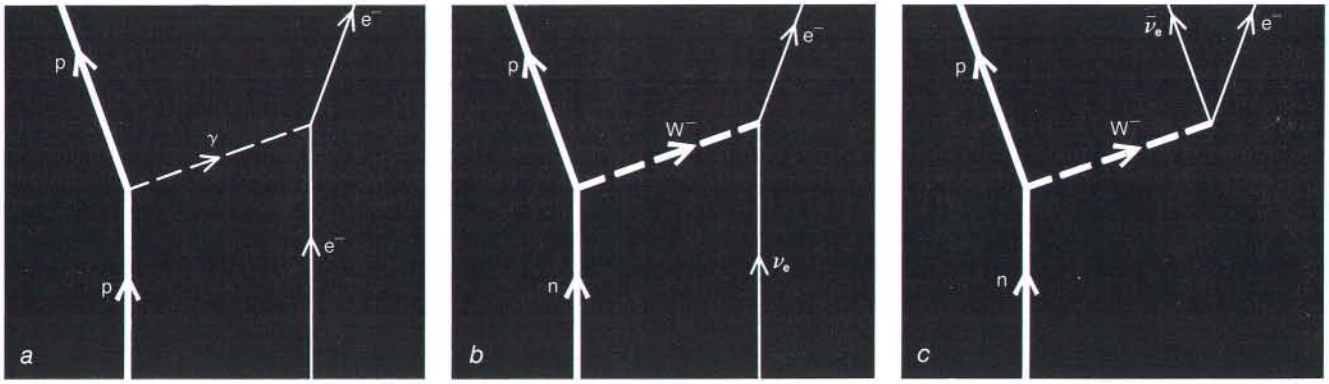
	GRAVITATIONAL	ELECTRO-MAGNETIC	STRONG	WEAK
RANGE	∞	∞	10^{-13}–10^{-14} CM.	$<<10^{-14}$ CM.
EXAMPLES	ASTRONOMICAL FORCES	ATOMIC FORCES	NUCLEAR FORCES	NUCLEAR BETA DECAY
STRENGTH (NATURAL UNITS)	$G_{NEWTON} = 5.9 \times 10^{-39}$	$e^2 = \frac{1}{137}$	$g^2 = 1$	$G_{FERMI} = 1.02 \times 10^{-5}$
PARTICLES ACTED UPON	EVERYTHING	CHARGED PARTICLES	HADRONS	HADRONS LEPTONS
PARTICLES EXCHANGED	GRAVITONS	PHOTONS	HADRONS	?

FOUR TYPES OF INTERACTION among particles are believed to account for all physical phenomena. "Range" is the distance beyond which the interaction effectively ceases to operate. In two cases, the range is believed to be infinite. "Strength" is a dimensionless number that characterizes the strength of the force under conditions typical of current observations. Thus, the gravitational force is some 39 orders of magnitude weaker than the strong force.

Quantum field theory was born in the late 1920s through the union of special relativity and quantum mechanics. It is easy to see how relativity leads naturally to the field concept. If I suddenly give one particle a push, my action cannot produce any instantaneous change in the forces (gravitational, electromagnetic, strong or weak) that are acting on a neighboring particle because according to relativity no signal can travel faster than the finite speed of light. In order to maintain the conservation of energy and momentum at every instant, we say that the pushed particle produces a field, which carries energy and momentum through surrounding space and eventually hands some of it over to the neighboring particle. When quantum mechanics is applied to the field, one discovers that the energy and momentum must come in discrete chunks, called quanta, which we identify with the elementary particles. Thus, relativity and quantum mechanics lead us naturally to a mathematical formalism, quantum field theory, in which elementary particle interactions are explained by the exchange of elementary particles themselves.

It is a simple consequence of the uncertainty principle in quantum mechanics (which states that the uncertainties in our knowledge of the momentum and the position of a particle are inversely proportional to each other) that the range of the force should be inversely proportional to the mass of the exchanged particle. (For an exchanged mass equal to that of the proton, the range is about 2×10^{-14} centimeter.) Thus, electromagnetism and gravitation, which seem to be of infinite range, are due to the exchange of particles of zero mass: the familiar photon and the hypothetical graviton. The strong interactions are generally believed to arise from the exchange of a large variety of strongly interacting particles, including protons, neutrons, mesons and hyperons of various kinds. Inasmuch as the weak interactions have a much shorter range than the strong interactions, they must be produced by the exchange of much heavier particles, presumably particles too heavy to have yet been created with existing accelerators.

For many years, it has been speculated that there may be a deep relation between the weak interactions and the electromagnetic interactions, with the difference in their apparent strengths



ELECTROMAGNETIC AND WEAK PROCESSES exhibit striking similarities when depicted in the form of Feynman diagrams. Such diagrams, developed by the physicist Richard P. Feynman, symbolize the interactions that underlie subnuclear phenomena, for example, the collision between two particles, which physicists refer to as a scattering event. Thus, diagram *a* indicates that the electromagnetic scattering of an electron by a proton results from the exchange of a photon, which transfers energy and momentum from one particle to another. Since time proceeds upward in these diagrams, the photon in *a* is traveling from the proton to the electron, but the diagram is intended also to cover the equally important case in which the photon is traveling in the opposite direction on a trajectory from lower right to upper left. It is precisely this feature of lumping together different processes in a single graph that constitutes the great conceptual value of the "language" of the Feynman diagrams. The scattering of a neutrino by a neutron (*b*) represents a weak interaction in which a heavy particle as yet undetected, the intermediate vector boson, $W^{\pm}$, is believed to play a role analogous to that of the photon in electromagnetic scattering. In this Feynman diagram the *W* particle is assigned a negative charge because it is assumed to be traveling from left to right. The particle could equally well be regarded as carrying a positive charge and traveling from right to left. The intermediate vector boson is also thought to mediate the radioactive decay of a neutron into a proton, an electron and an antineutrino (*c*). Note that diagram *b* can be obtained from diagram *a* simply by changing some of the particles into others of different charge. Diagram *c* can be obtained from diagram *b* by replacing the incoming neutrino by an outgoing antineutrino. It is also to be noted that the total charge is conserved at each vertex in the diagrams.

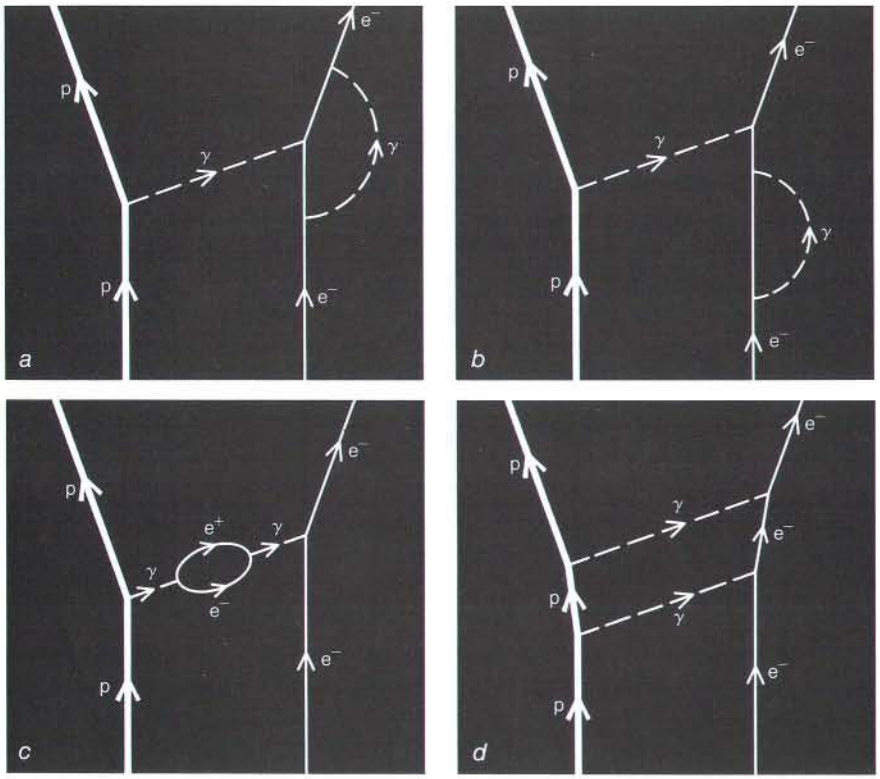
being due simply to the large mass of the particle exchanged in the weak interactions. This hypothesis is supported by the observation that the angular momentum exchanged in weak processes such as nuclear beta decay has the same value as the angular momentum of a single photon (equal to the Planck constant: 1.0546×10^{-27} erg-second). In fact, the hypothetical particle, or quantum, exchanged in weak interactions has long had a name: the intermediate vector boson. (The term "vector" is used because any particle with this angular momentum is usually described by a field that is a four-dimensional vector, like the vector potential used to describe the photon in James Clerk Maxwell's theory of electromagnetism. The term "boson" refers to the entire class of particles whose angular momentum is an integer multiple of Planck's constant.)

If we assume that the intrinsic interaction strength of the putative intermediate vector boson that is exchanged in the weak interactions is the same as it is for a photon, then the weak force will have the same strength as the electromagnetic force at short distances; it only appears weaker because of its much shorter range. The effect of the weak force is reduced in any given process by a factor given by the square of the ratio of the typical masses involved in the process to the mass of the intermediate vector boson. For processes characterized by masses comparable to the mass of the proton, the weak force is roughly 1,000 times weaker than the electromagnetic force. Hence, taking the square root, we conclude that the mass of the intermediate vector boson is roughly 30 proton masses. By applying the conservation of charge to processes such as nuclear beta decay, we also see that the intermediate vector boson must carry either a positive or a negative electric charge equal in magnitude to the charge of the proton or of the electron.

A quantum field theory tells us how to calculate the rate for any process in terms of a sum of individual processes, each symbolized by a Feynman diagram such as the one shown above and elsewhere in this article. This useful method for visualizing subnuclear events was introduced some 25 years ago by Richard P. Feynman and led to a solution of the one great problem that had plagued quantum field theory almost from its birth: the problem of infinities. It was found in the 1930s that the contributions produced by processes more complicated than single-particle exchanges usually turn out to be infinitely large. In fact, the electrostatic repulsion within a single electron produces an infinite self-energy, which manifests itself whenever a photon is emitted and reabsorbed by the same electron [*see illustration on next page*]. These infinities arise only in Feynman diagrams with loops, and they can be traced to the infinite number of ways that energy and momentum can flow through the loop from one particle to another.

As is usually the case when paradoxes arise in science, the problem of infinities is simultaneously a curse and a blessing. It is a curse because it keeps us from getting on with calculations we would like to carry out. It is a blessing because when the solution is found, it may work only for a limited class of theories, among which one hopes to find the true theory.

That is just what seems to have happened with the problem of infinities. In the late 1940s a group of young theoreticians working independently (Feynman, then at Cornell University, Julian Schwinger at Harvard University, Freeman J. Dyson at the Institute for Advanced Study and Sin-itiro Tomonaga in Japan) found that in a certain limited



HIGHER-ORDER CONTRIBUTIONS TO SCATTERING RATES were formerly impossible to calculate when a photon was emitted and reabsorbed by the same electron (*a, b*) or when a photon gave rise to an electron-positron pair that subsequently recombined into a photon (*c*). When physicists see "loops" of this kind in Feynman diagrams, they are prepared to encounter infinities in trying to calculate reaction rates. Thus, in diagram *b* the electrostatic repulsion within a single electron manifests itself as an infinite self-energy. It was found in the late 1940s that such infinities can be handled by redefining the mass and charge of the electron, a process called renormalization. The infinity problem does not arise, however, in scattering events such as that in *d*, where the loop has four corners or more.

class of field theories the infinities occur only as "renormalizations," or corrections, of the fundamental parameters of the theory (such as masses and charges) and can therefore be eliminated if one identifies the renormalized parameters with the measured values listed in tables of the fundamental constants. For example, the measured mass of the electron is the sum of its "bare" mass and the mass associated with its electromagnetic self-energy. In order for the measured mass to be finite, the bare mass must have a negative infinity that cancels the positive infinity in the self-energy. One simple version of the field theory of electromagnetic interactions not only was found to be renormalizable in the sense that all infinities could be eliminated by a renormalization of the electron's mass and charge but also led to electrodynamic calculations whose agreement with experiment is without precedent in physical science. Thus the theory predicts that the value of the magnetic moment of the electron (in natural units) is 1.0011596553, whereas the observed value is 1.0011596577. The uncertainty in both figures is in the ninth place: +0.0000000030.

In spite of this stunning success, attempts to construct renormalizable field theories of the other elementary particle interactions long proved to be unsuccessful. For the strong interactions, there was no lack of possible renormalizable theories; rather the trouble was (and indeed still is) that the strength of the interaction invalidates any simple approximation scheme that might be used to draw consequences from a given field theory that could be checked by experiment. (Roughly speaking, the probability of exchanging a set of strongly interacting particles in a high-energy collision is independent of the number of particles exchanged, with the result that extremely complicated exchanges have to be taken into account in even the lowest approximation.)

For the gravitational interactions, we have a well-known field theory, Einstein's general theory of relativity, which accounts very well for phenomena on the scale of the solar system but seems not to be renormalizable and presumably therefore needs modification for phenomena at very short distances. The problem in this instance is the opposite of that for the strong interactions: gravitational effects are so weak that one can get no help from experimental measurements, at their current level of precision, in finding the correct theory.

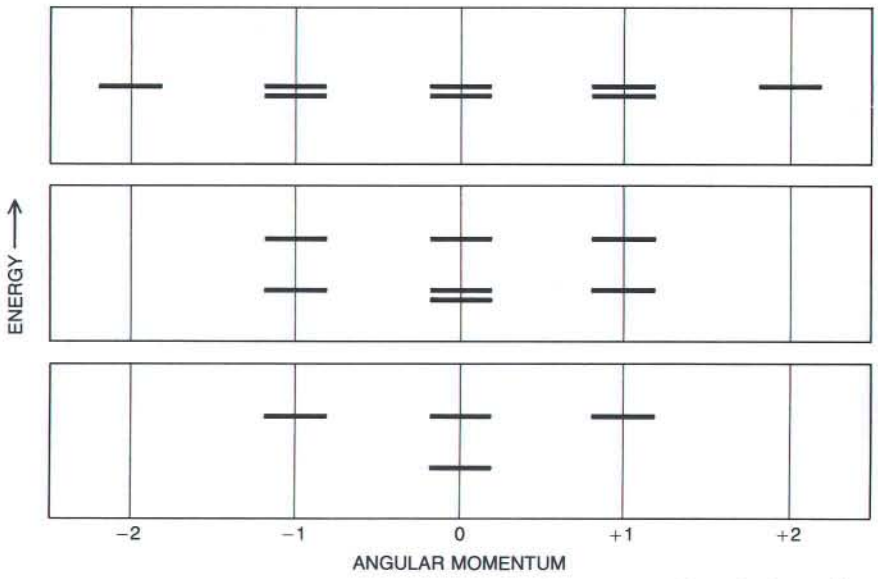
The weak interactions present an intermediate case: they are strong enough so that good experimental data are available (although nowhere near as copious as for the strong interactions) and yet weak enough so that approximate calculations are practicable. Even though the weak interactions are believed to be similar to the electromagnetic interactions, however, the theory, as it existed until a few years ago, does not appear to be renormalizable. To put the problem more specifically, the exchange of pairs of intermediate vector bosons in processes such as neutron-neutrino scattering [*see illustration on page 29*] leads to infinities that cannot be absorbed in a renormalization of the parameters of the theory. For this reason, although the quantum field theory of intermediate vector bosons gave a perfectly good approximate picture of the observed weak interactions, it broke down as soon as it was pushed beyond the lowest approximation.

What is the difference between photons and intermediate vector bosons that makes the infinities so much worse for the latter? A detailed analysis enables us to trace the difference back to the fact that the photon has zero mass, whereas the intermediate vector boson has ponderable mass. Like all other zero-mass particles, the photon can exist as a superposition of at most two pure states, characterized by left or right circular polarization, in which the axis of rotation is respectively in the same direction as the direction of motion or in the opposite direction. On the other hand, the intermediate vector boson, like any other massy particle whose angular momentum equals Planck's constant, can exist in any one

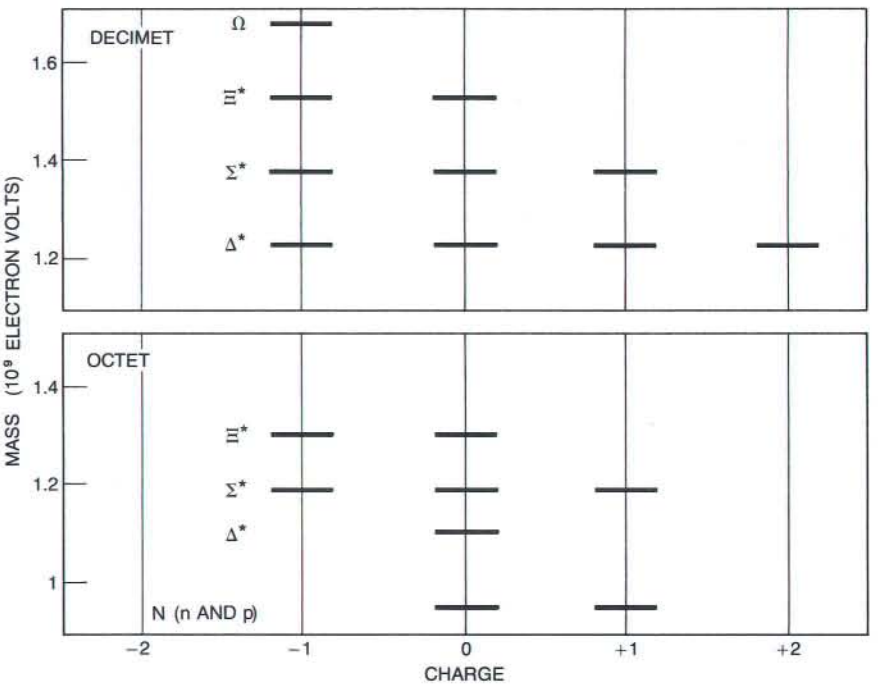
of three states, characterized by an axis of rotation that points in the direction of motion, points in the opposite direction or points in a direction perpendicular to the direction of motion. It is the exchange of intermediate vector bosons whose axes of rotation are perpendicular to the direction of motion that produces the nonrenormalizable infinities.

Before we can see how this problem can be resolved, we must first consider the role that symmetry principles have come to play in theoretical physics. Considerations of symmetry have always been important in science, but they acquired special significance with the advent of quantum mechanics. That is because the energy or mass levels of any quantum-mechanical system subject to a symmetry principle are generally required to form certain well-defined and easily recognized families. (In mathematical language, one says that the collection of all the mathematical operations on fields that leave the form of the field equations unchanged constitute a "group"; the levels with a given energy or mass are said to form a representation of that group.)

For example, the quantum-mechanical equations that describe the hydrogen atom obey the symmetry principle that all spatial directions are equivalent. As a result, the energy levels of hydrogen form families with an odd number of members (1, 3, 5 and so on), the levels within each family being distinguished by the orientation of the axis of rotation [*see top illustration at right*]. When we look at a table showing the masses of elementary particles, we find a similar grouping into families: the proton and the neutron have nearly equal mass; the three sigma hyperons (Σ^+, Σ^0, Σ^-) likewise have nearly the same mass, and so on. For this reason, it is believed that the field equations of elementary particle physics obey isotopic spin symmetry, a symmetry principle analogous to rotational symmetry except that the "rotation" alters the value of the particle's electric charge rather than its spatial orientation. In the early 1960s it was further realized that the various particle pairs, triplets and so on are themselves grouped into larger superfamilies of eight, 10 or even more members, reflecting an approximate symmetry larger than isotopic spin symmetry [*see bottom illustration at right*]. All these symmetry principles require that the field equa-



SYMMETRY PRINCIPLES require the energy levels of an atom, such as the lower levels of hydrogen shown here, to cluster in well-defined families. Each quantum state of the hydrogen atom is indicated by a short bar. The value of the energy is indicated schematically by the bar's vertical position. (The breaks between panels indicate energy gaps.) The value of the angular momentum around any fixed direction (in units of Planck's constant) is indicated by the horizontal position of the bar. The exact equality of energies within the various triplets, quintuplets and so on is a consequence of the rotational symmetry of the equations that describe the atom, whereas the approximate equality of energies within the various larger families is dictated by more detailed features of the dynamics, such as the weakness of the magnetic coupling between proton and electron and the small value of the electron's charge.



FAMILIES OF ELEMENTARY PARTICLES are believed to be a consequence of a symmetry principle known as isotopic spin symmetry, which is analogous to the rotational symmetry that produces the families of quantum states within the hydrogen atom. The grouping of these families of elementary particles into superfamilies (octets, decimets and so on) was proposed independently in the early 1960s by Murray Gell-Mann and Yuval Ne'eman. The vertical position of the bars indicates the mass of the particles. Their horizontal position indicates the electric charge in units of the proton's charge. Particles marked with an asterisk are very short-lived states.

tions do not change when we simultaneously perform well-defined "rotations" on some labeled characteristic of a family or superfamily of particles everywhere in space.

One can imagine a much more powerful requirement: that the equations should not change when we perform such rotations of labeled characteristics independently at each point in space and time. The first kind of symmetry operation—the less stringent one—is comparable to giving each apple in a basket the same rotation from one orientation in space to another. The second and more general kind of symmetry operation is comparable to rotating each apple in the basket separately to different new orientations. Invariance under the second kind of symmetry operation is known as a gauge symmetry.

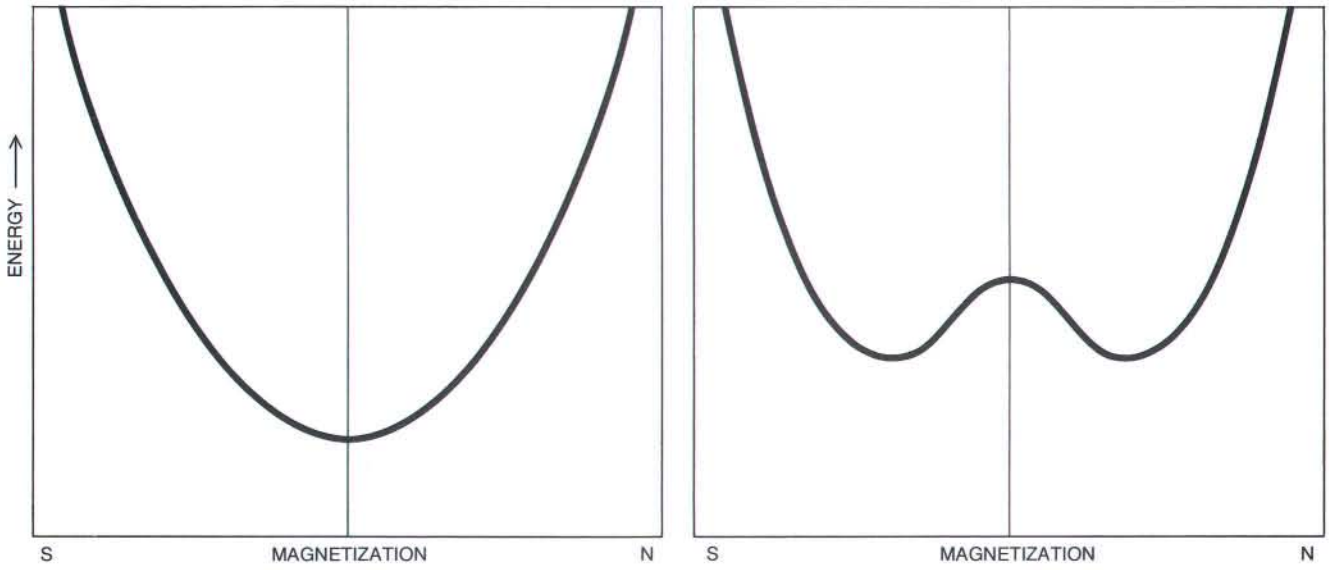
It has been known for many years that the Maxwell field equations of electromagnetism obey a gauge symmetry, based on the group of rotations in two rather than in three directions. Indeed, the logic can be turned backward: assuming this gauge-symmetry principle, one can deduce all the properties of electromagnetism, including Maxwell's equations and the fact that the mass of the photon is zero. It is difficult to conceive of any better example of the power of symmetry principles in physics.

Our hopes of perceiving an underlying identity in the weak and the electromagnetic interactions lead us naturally to suppose there may be some larger gauge symmetry that forces the photon and the intermediate vector boson into a single family. (Indeed, the mathematical theory of generalized gauge symmetries has been understood since the work in 1954 of C. N. Yang and Robert L. Mills, who were then working at Brookhaven National Laboratory.) For this to be possible, however, the intermediate vector boson, like the photon, would have to have zero mass, and we have already seen that its mass is actually much greater than that of any known particle. How can there be any family connection between two such different particles?

The answer to this conundrum lies in considerations of appearance and reality. Inasmuch as symmetry principles govern the form of the field equations, they are generally regarded as providing information about the laws of nature on the deepest possible level. Is it conceivable for a symmetry principle to be valid on this level and yet not be manifest in the masses and other observed properties of physical particles? The familiar phenomenon of ferromagnetism provides an example of how this can happen.

The equations governing the electrons and iron nuclei in a bar of iron obey rotational symmetry, so that the free energy of the bar is the same whether one end is made the north pole by magnetization or the south. At high temperatures the curve of energy versus magnetization has a simple U shape that has the same rotational symmetry as the underlying equations [*see illustration below*]. The equilibrium state, which is the state of lowest energy at the bottom of the U, is also a state of zero magnetization, which shares this symmetry.

On the other hand, when the temperature is lowered, the lowest point on the U-shaped curve humps upward so that the curve resembles a W with rounded corners. The curve still has the same rotational symmetry as the underlying equations, but now the equilibrium state has a definite nonzero magnetization, which can be either north or south but which in either case no longer exhibits the rotational symmetry of the equations. We say in such cases that the symmetry is spontaneously broken. A tiny physicist liv-



EXAMPLE OF "BROKEN" SYMMETRY can be found in the two different curves that result when one plots the free energy versus magnetization for a bar magnet at high temperature (*left*) or at low temperature (*right*). The magnet naturally seeks a state of minimum free energy. At high temperature this is a state of zero magnetization, a state that exhibits perfect symmetry between north and south. At low temperature the equilibrium state shifts to one of nonzero magnetization, which can be either north or south, even though the free-energy curve is still perfectly symmetric between north and south. In a situation of this kind, physicists say that the symmetry is spontaneously broken. The author invokes the hypothesis of a similar breaking of symmetry to unify the electromagnetic and weak interactions.

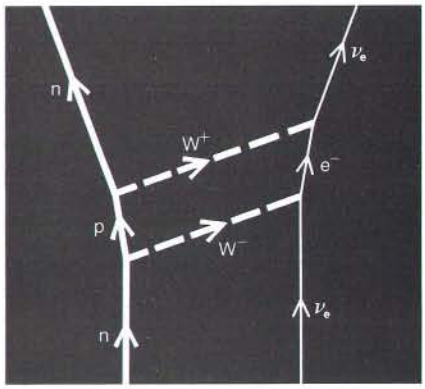
ing inside the magnet might not even know that the equations of the system have an underlying rotational symmetry, although we, with our superior perspective, find this easy to recognize. Reasoning by analogy, we reach the conclusion that a symmetry principle might thus be exactly true in a fundamental sense and yet not be visible at all in a table of elementary-particle masses.

The first proposed example of a broken symmetry of this kind in elementary particle physics was a nongauge symmetry known as chiral symmetry. (The term "chiral" is from the Greek for "hand," and it is employed in this case because the symmetry consists of independent three-dimensional rotations on fields of left-handed or right-handed polarization. This symmetry contains within it the unbroken three-dimensional rotation group of isotopic spin symmetry.) Chiral symmetry has led to great successes in predicting the properties of low-energy pi mesons, but a discussion of such matters would take us too far afield.

In 1967 I suggested that the weak and electromagnetic interactions are governed by a broken gauge-symmetry group. (A similar suggestion was made independently some months later by Abdus Salam of the International Center for Theoretical Physics in Trieste.) The proposed group contains within it the unbroken gauge-symmetry group of electromagnetism and therefore requires the photon to have zero mass, but the other members of the photon's family are associated with broken symmetries and therefore pick up a large mass from the symmetry breaking. In the simplest version of this theory the relatives of the photon would consist of a charged intermediate vector boson (long referred to as the W particle) with a mass greater than 39.8 proton masses plus an additional neutral intermediate vector boson (which I called the Z particle) with a mass greater than 79.6 proton masses. (A theory of this kind is more closely analogous to superconductivity than to ferromagnetism: in a superconductor, electromagnetic gauge symmetry is broken and the photon itself acquires mass, as is shown by the fact that a magnetic field can penetrate only a short distance into a superconductor. In particle physics the appearance of vector boson masses in this way is called the Higgs mechanism because it first became known as a mathematical possibility through a paper written in 1964 by Peter Higgs of the University of Edinburgh.)

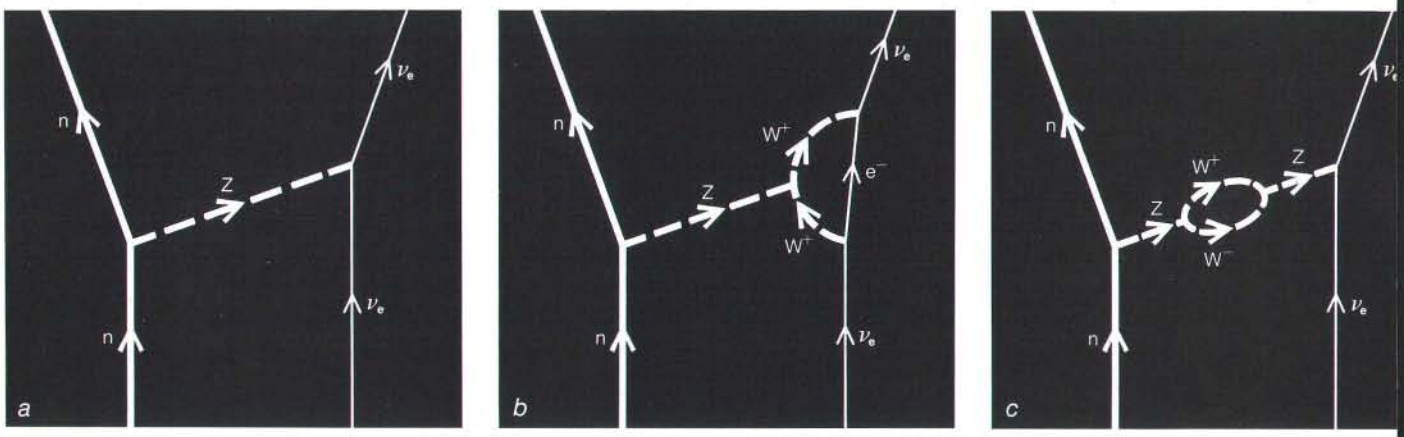
At the time I proposed my theory, there was no experimental evidence for or against it and no immediate prospect of getting any. There was, however, an internal test of the theory that could be made without help from experiment. We have seen that the infinities in the quantum field theory of pure electromagnetism can be renormalized away, whereas this cannot be done with the existing theory of weak interactions in which intermediate vector bosons have mass. Thus, one can ask: Does a field theory become renormalizable when the intermediate vector bosons belong to the same family as the photon and acquire mass only through the spontaneous breakdown of a gauge symmetry? I had suggested in 1967 that this might be the case, but the renormalizability of the theory was not demonstrated until four years later, when it was first shown by Gerhard 't Hooft, then a graduate student at the University of Utrecht. (The proof of this point has since been made more rigorous through the work of a number of theorists, particularly B. W. Lee, J. Zinn-Justin, M. Veltman and 't Hooft himself.) It turns out that the various multiparticle exchanges involving photons, charged intermediate vector bosons, neutral intermediate vector bosons and other particles add up so as to cancel all nonrenormalizable infinities [*see top illustration on next two pages*].

Once the renormalizability of the theory was established, it became clear that the long-sought goal of a unified field theory of weak and electromagnetic interactions might finally be at hand. It then became crucially important to test the theory against experiment. Until such time as intermediate vector bosons can be produced directly, the best way to test the theory is to look for effects attributable to the newly predicted neutral intermediate vector boson—the new Z particle—that must appear in the same family as the photon and the charged intermediate vector bosons. The neutral boson does not contribute to processes such as beta decay, in which a charge must be exchanged between the nucleus and the emitted particles. It does, however, contribute along with the charged bosons in processes such as the scattering of "ordinary" neutrinos by electrons [*see "a" and "b" in bottom illustration on next two pages*] and materially changes their rate. Finally, there are processes such as the elastic scattering of "muon-type" neutrinos by electrons and the elastic scattering of any type of neutrino by protons or neutrons [*see "c" and "d" in bottom illustration on next two pages*] that could be produced only by exchanges of neutral intermediate vector bosons.



INFINITY PROBLEM ARISES in calculating the rate of neutron-neutrino scattering events involving the sequential (and hypothetical) exchange of two intermediate vector bosons, W^- and W^+. The first exchange converts the neutron into a proton and the neutrino into an electron. The second exchange restores the original cast of characters. Note that the Feynman diagram of this neutron-neutrino event has the same form as diagram *d* in the illustration on page 26, in which the infinity problem does not arise. The author's recent work, however, indicates how infinities can be removed in processes involving the intermediate vector boson.

For some years, these neutral-current processes, as they are called, remained at the edge of detectability, and many physicists doubted their existence. Within the past year, however, evidence for neutral-current processes has at last begun to appear. A pan-European collaboration involving some 55 investigators from seven different institutions, working at the European laboratory for particle physics (CERN) near Geneva, has found two events in which muon-type antineutrinos are scattered by electrons and several hundred events in which they are scattered by protons or neutrons. Such scattering events can apparently be explained only by the exchange of a neutral intermediate vector boson, or Z particle, and are therefore direct evidence for a new kind of weak interaction. Moreover, the inferred collision



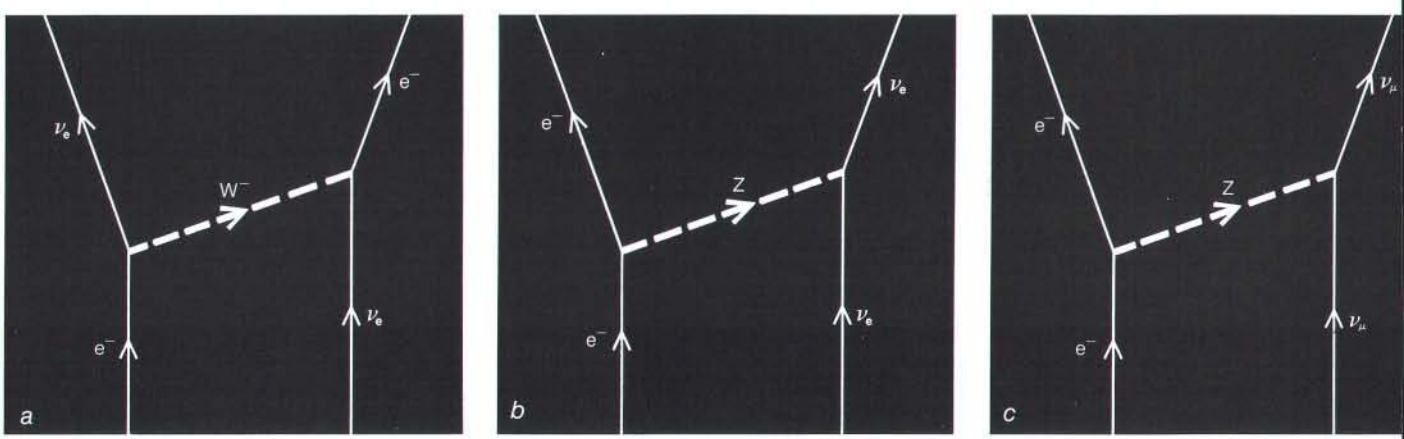
POSSIBLE SOLUTION OF INFINITY PROBLEM in calculating rates of neutron-neutrino scattering may be achieved by postulating the existence of the *Z* particle, a neutral intermediate vector boson predicted by unified theory of weak and electromagnetic interactions proposed by the author. The *Z* particle should lead to a variety of neutrino-scattering events such as these. When such processes are added to those involving the charged intermediate vector boson, $W^{\pm}$, the infinities appear-

rates agree well with rates predicted by the new theory. An American consortium working at Fermi National Accelerator Laboratory in Batavia, Ill., and another group working at Argonne National Laboratory have apparently also found neutral-current events [*see illustration on page 23*]. Further experiments aimed at the detection of neutral-current processes and aimed in addition at the measurement of their rates are in train at various laboratories in Western Europe, the U.S. and the U.S.S.R.

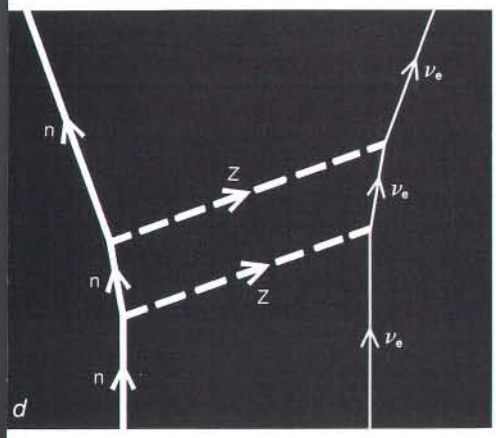
The existence of neutral-current processes is not yet definitely established, and in any case the general idea of a renormalizable unified field theory based on a spontaneously broken gauge symmetry does not depend absolutely on the existence of neutral-current processes. For instance, in one model suggested by Howard Georgi and Sheldon L. Glashow of Harvard University, the photon and the charged intermediate vector boson form a family by themselves, although this simplification is achieved at the cost of introducing new particles of other kinds. (There are now a host of other ingenious models suggested by more theorists than can be named here.) There is no doubt, however, that the apparent detection of neutral-current processes has brought welcome encouragement to field theorists.

At the same time that experimentalists have been working to test the consequences of unified weak and electromagnetic field theories, theoreticians have been discovering that the new theories freshly illuminate a number of outstanding problems. One, for instance, has to do with the dynamics of the giant stellar explosions known as supernovas. It is believed that supernovas occur at a certain point in the life of a very massive star when the core of the star becomes unstable and begins to implode, or collapse. It has long been a puzzle how the implosion can be reversed and become an explosion, and how the star can shed enough of its outer layers to reach stability as an ultradense neutron star, only 10 or 20 kilometers in diameter. (There is observational evidence that at least two pulsars, believed to be rapidly rotating neutron stars, are embedded in the remnants of past supernovas.)

In 1966 Stirling A. Colgate and R. H. White of the New Mexico Institute of Mining and Technology suggested that the outer layers of an exploding star

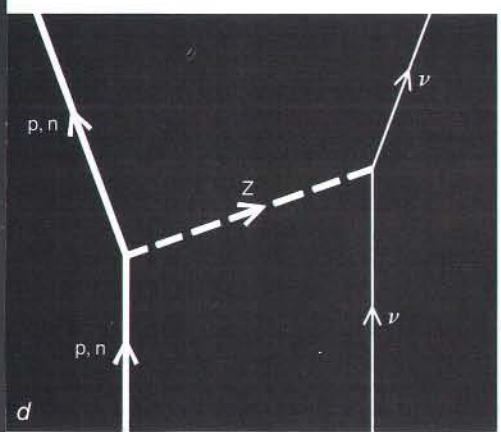


TESTS FOR EXISTENCE OF *Z* PARTICLE can be made by studying interactions of neutrinos with electrons or with protons and neutrons. The scattering of an electron-type neutrino, ν_e, by an electron can involve the exchange of either a charged intermediate vector boson $W^{\pm}$ (*a*) or a neutral intermediate vector boson *Z* (*b*). Hence, the process can be used to test for the *Z* particle only if the rate of the process is carefully measured and compared with theory. In contrast, scattering of a muon-type neutrino, ν_μ, by an electron (*c*) or of any kind of neutrino by a proton or neutron (*d*) can occur only by ex-



ing in the sum of the contributions to the total rate, symbolized by all such diagrams, can be absorbed into a renormalization of parameters of new theory.

might be blown off by the pressure of neutrinos produced in the hot stellar core, but detailed calculations by James Wilson of the Lawrence Livermore Laboratory of the University of California, using the then available theories of weak interactions, did not support the conjecture. It has recently been pointed out by Daniel Freedman of the State University of New York at Stony Brook that neutral currents can produce "coherent" neutrino interactions, in which a neutrino interacts with an entire nucleus rather than with its individual neutrons and protons. This interaction leads in turn to a much stronger interaction between neutrinos and the relatively heavy nuclei, chiefly the nuclei of iron, in the outer layers of the stellar core. According to the latest calculations made by Wilson, the increased neutrino pressure is apparently sufficient to produce a supernova.



change of a *Z* particle and therefore provides direct evidence for a new kind of weak interaction. These interactions are neutral-current processes as shown at the top right on page 23.

Another old problem that may be solved through the development of unified gauge theories of weak and electromagnetic interactions concerns the origin of the slight departures from perfect isotopic spin symmetry. The masses of particles within a given family are not precisely equal, generally differing by less than 1 percent to several percent. (The masses of the best-known family pair, the neutron and the proton, differ by only 0.13 percent.) The differences in mass are about what one would expect if isotopic spin symmetry were respected by the strong interactions but violated by the electromagnetic ones. Calculations along these lines, however, never seem to work. For instance, the electromagnetic self-energy of the proton not only turns out to be positive, contrary to the observation that the neutron is slightly heavier than the proton, but also has an infinite value. This infinity is of the type discussed above, but it cannot be eliminated by renormalization of the bare mass of the proton, if we insist that the bare masses of the proton and the neutron are equal.

If, as now seems possible, the weak interactions really have an intrinsic strength comparable to that of the electromagnetic interactions, they can provide additional corrections to isotopic spin symmetry that can cancel the infinities due to electromagnetism and leave a finite correction of the right magnitude and sign. Before such calculations can be effectively carried out, however, it is necessary to settle on a detailed model not only of the weak interaction of electrons and neutrinos, as in the 1967 theory, but also of the weak interactions of the strongly interacting particles. This task is still in progress.

Since the weak and electromagnetic interactions seem to be described by a unified gauge-symmetric field theory, it is natural to ask whether the strong interaction can be brought into this picture. There are in fact good reasons for seeking a description of strong interactions in terms of gauge field theories. Possibly the most important is that for a certain class of such theories it is possible to prove that the strong and electromagnetic interactions must necessarily exhibit symmetries between right and left and between matter and antimatter, as is in fact observed to be the case, even though these symmetries are not respected by the weak interactions. As we have seen, the difficulty in testing such field theories is not the lack of experimental data but rather the lack of a method of calculation that can cope with the strength of the strong interactions. Within the past year, however, there has been a theoretical breakthrough that may at last make possible a solution of this problem. David Politzer, a graduate student at Harvard, and independently David Gross and Frank Wilczek of Princeton University have discovered that in certain gauge field theories the effective strength of the strong interactions at a given energy decreases as the energy rises. In such "asymptotically free" theories it is possible to carry out approximate calculations with the same methods one uses for the weak and electromagnetic interactions, provided that one works at an energy sufficiently high (no one really knows how high) for the strong interactions to be sufficiently weak. Some of the calculations carried out in this way seem to agree quite well with experiment, whereas other calculations do not.

Although it is too early to tell how this will all work out, the development of asymptotically free gauge theories has already led Gross, Wilczek, Politzer, Georgi, Glashow, Helen Quinn and me to an intriguing series of conjectures. If the effective interaction strength becomes small at high energies and short distances, then it must become large at low energies and large distances. Perhaps this conjecture explains why the ordinary elementary particles cannot be broken up into quarks: as a quark is pulled away from the rest of the particle, the forces may increase without limit. Perhaps the intrinsic interaction strength of the strong interactions is really of the same order of magnitude as that of the weak and electromagnetic interactions and only appears stronger because our present experiments happen to be carried out at relatively low energies and large distances. Perhaps the strong interactions are really caused by the exchange of particles that belong to the same family as the photon and the intermediate vector bosons that are responsible for the electromagnetic and weak interactions. If these speculations prove to be borne out by further theoretical and experimental work, we shall have moved a long way toward a unified view of nature.

The Number of Families of Matter

How experiments at CERN and SLAC, using electron-positron collisions, showed that there are only three families of fundamental particles in the universe

by Gary J. Feldman and Jack Steinberger

The universe around us consists of three fundamental particles. They are the "up" quark, the "down" quark and the electron. Stars, planets, molecules, atoms—and indeed, ourselves—are built from amalgamations of these three entities. They, together with the neutral and possibly massless partner of the electron, the electron neutrino, constitute the first family of matter.

Nature, however, is not so simple. It provides two other families that are like the first in every respect except in their mass. Why did nature happen to provide three replications of the same pattern of matter? We do not know. Our theories as yet give no indication. Could there be more than three families? Recent experiments have led to the conclusion that there are not.

GARY J. FELDMAN and JACK STEINBERGER were leaders in the effort to determine experimentally the number of families of matter. Feldman received his doctorate from Harvard University in 1971 and spent the following 19 years studying electron-positron annihilation at the Stanford Linear Accelerator Center (SLAC). He was co-leader of the Mark II, the experimental facility of the Stanford Linear Collider (SLC). Last fall he moved to Harvard and began studying proton-antiproton collisions at the Fermilab National Accelerator Laboratory in Batavia, Ill. Steinberger was born in Germany, came to the U.S. as a child and received his doctorate from the University of Chicago in 1948. Since 1968 he has been associated with the European laboratory for particle physics (CERN) near Geneva. Between 1983 and 1990 he headed Aleph, one of four experiments installed at the organization's Large Electron-Positron (LEP) Collider. He shared the 1988 Nobel Prize in Physics for his discovery of the muon neutrino in 1962.

In the spring and summer of 1989, experiments were performed by teams of physicists working at the Stanford Linear Accelerator Center (SLAC) and the European laboratory for particle physics (CERN) near Geneva. The teams used machines of differing designs to cause electrons (e^-) and positrons (e^+) to collide and thus produce quantities of the Z particle (or Z^0, pronounced "zee zero" or "zee naught").

The most massive elementary particle observed, the Z weighs about 100 times as much as a proton and nearly as much as an atom of silver. As we shall see, this mass is merely an average. The Z lifetime is so short that individual Z particles differ slightly in their mass. The spread in the mass values is called a mass width, a quantity that depends on the number of families of matter. Because this width can be measured experimentally, the number of families of matter can be inferred. In this article we describe the experiments by which the families of matter were numbered.

But let us first put this achievement into perspective. The past two and a half decades have witnessed a remarkable systematization of our knowledge of the elementary particles and their interactions with one another. The known particles can be classified either as fermions or as gauge bosons. Fermions are particles of spin 1/2, that is, they have an intrinsic angular momentum of $1/2\ \hbar$, where $\hbar$ is the Planck unit of action, 10^{-27} erg-second. Fermions may be thought of as the constituents of matter. Gauge bosons are particles of spin 1, or angular momentum $1\hbar$. They can be visualized as the mediators of the forces between the fermions. In addition to their spins, these particles are characterized by their masses and by their various couplings with one another, such as electric charges.

All known couplings, or interactions, can be classified into three types: electromagnetic, weak and strong. (A fourth interaction, gravity, is negligible at the level of elementary particles, so it need not be considered here.) Although the three interactions appear to be different, their mathematical formulation is quite similar. They are all described by theories in which fermions interact by exchanging gauge bosons.

The electromagnetic interaction, as seen in the binding of electrons and nuclei to form atoms, is mediated by the exchange of photons—the electromagnetic gauge bosons. The weak interaction is mediated by the heavy W^+, W^- and Z bosons, whereas the strong interaction is mediated by the eight massless "gluons." The proton, for instance, is composed of three fermion quarks that are bound together by the exchange of gluons.

These interactions also describe the creation of particles in high-energy collisions. The conversion of a photon into an electron and a positron serves

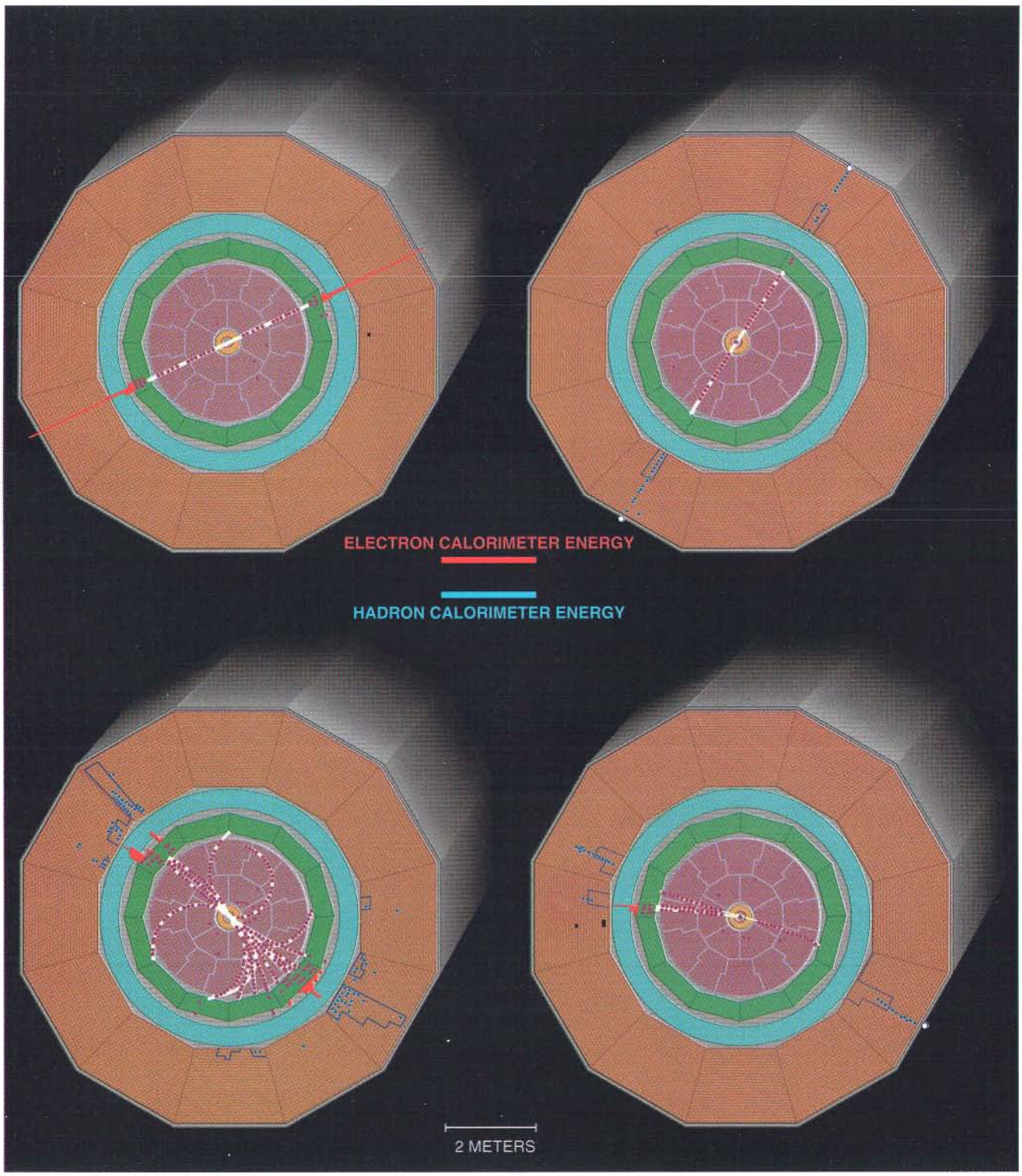
ALEPH DETECTOR, one of four at the Large Electron-Positron (LEP) Collider at CERN near Geneva, recorded these typical decays of Z particles. The cross-sectional diagrams show Z decay products as they traverse the detector. The four decays are (*clockwise, from upper left*) an electron and a positron, which appear as a single line of dots; two muons, which match the electrons' paths but penetrate the outer tracking devices; two tau leptons, one of which has decayed into a muon and two unseen neutrinos, accompanied by another that has decayed into three pions; and two quarks, which form hadron jets. Most Zs decay to quarks. Histograms (*blue and red*) represent particle energies.

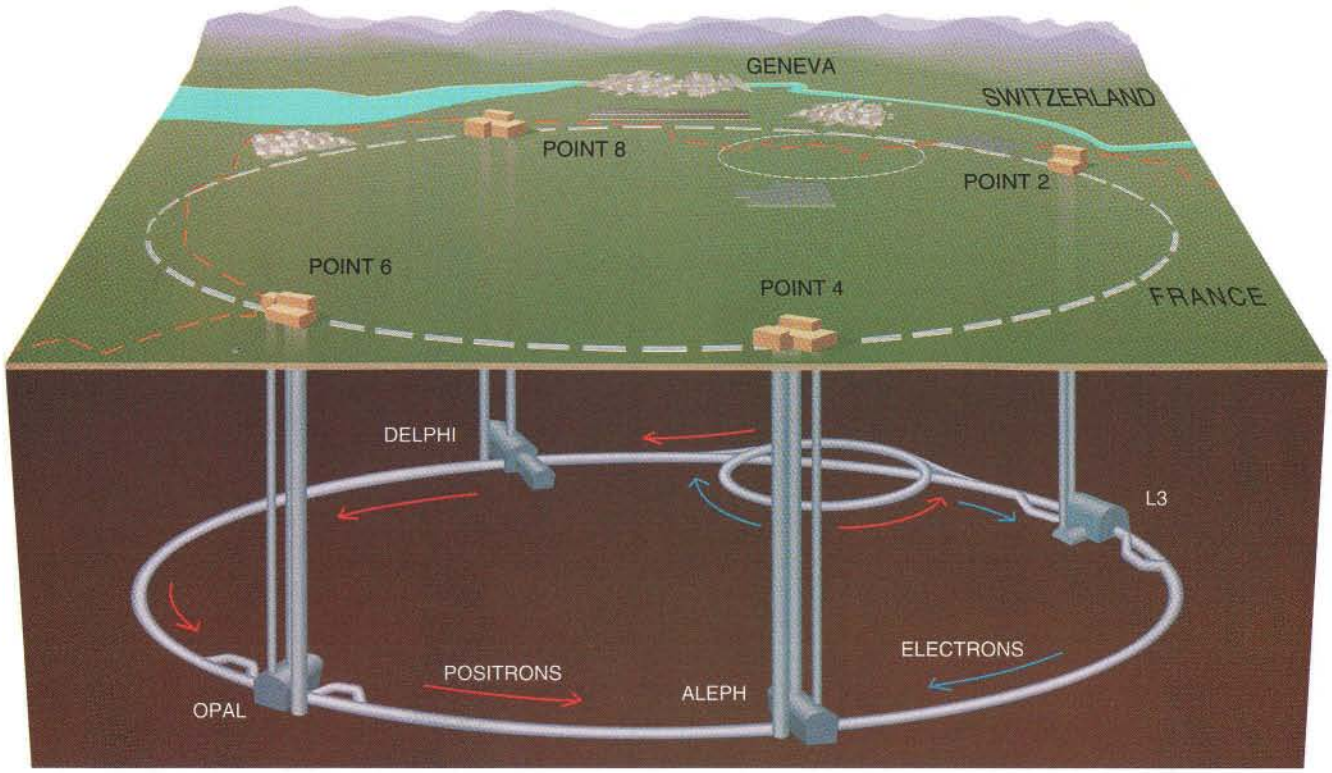
Reprinted from the February 1991 issue

as an example. So does the annihilation of an electron colliding with a positron at immensely high energy to produce a Z particle.

The evolution of these gauge theories constitutes a strikingly beautiful advance in particle physics. The unification of electromagnetism with the weak interaction was put forward during the years 1968–1971. This "electroweak" theory predicted the neutral weak interaction, discovered at CERN in 1973, and the heavy intermediate bosons W^+, W^- and Z^0, discovered 10 years later, also at CERN [see "Unified Theories of Elementary-Particle Interaction," by Steven Weinberg; SCIENTIFIC AMERICAN, July 1974].

The gauge theory of the strong interaction was advanced in the early 1970s. This theory is called quantum chromodynamics because it explains the strong force by which quarks interact on the basis of their "color." Despite its name, color is an invisible trait. It is to the strong interaction what charge is to the electrical one: a quantity that characterizes the force. But whereas electrodynamic charge has only one state—positive or negative—the color charge has three. Quarks come in red, green and blue; antiquarks come in antired, antigreen and antiblue.





LARGE ELECTRON-POSITRON COLLIDER creates *Z* bosons by bringing electrons and positrons into collision in a storage ring 27 kilometers in circumference. The particles countercirculate in bunches. Magnets confine the two beams to their proper orbits, and radio-frequency power accelerates them to a combined energy near 90 billion electron volts, equivalent to the *Z* mass. The bunches meet head-on 45,000 times a second at points inside the Aleph, Opal, Delphi and L3 detectors.

Together these two gauge theories predict, often with quite high precision, all elementary phenomena that have so far been observed. But their apparent comprehensiveness does not mean that the model is complete and that we can all go home. Gauge theory predicts the existence of the so-called Higgs particle, which is supposed to explain the origin of particle mass. No physicist can be happy until it is spotted or a substitute for it is supplied. Gauge theory also includes a number of arbitrary physical constants, such as the coupling strengths of the interactions and the masses of the particles. A complete theory would explain why these particular values are found in nature.

Among the rules the electroweak theory does provide is one that requires fermions to come in pairs. The electron and electron neutrino are such a pair; they are called leptons because they are relatively light. Another rule is that each particle must have its antiparticle—against the electron is posed the positron; against the electron neutrino, the electron antineutrino. When particles and antiparticles collide, they can annihilate one another, producing secondary particles. Such reactions, as we shall see, underlie the experiments discussed here.

To avoid some subtle disasters in the theory, it is necessary to associate with a lepton pair a corresponding pair of quarks. The electron is the lightest charged lepton, and therefore it is associated with the lightest quarks, the *u* quark (or up quark) and the *d* quark (or down quark). Quarks have not been seen in the free state; they are only found bound to other quarks and antiquarks.

The proton, for example, is composed of two *u* quarks and a *d* quark, whereas the neutron is composed of two *d* quarks and a *u* quark. A complete second family and most of a third have been shown to exist in high-energy experiments. In each case, the particles are much more massive than the corresponding members of the preceding family (the neutrinos form a possible exception). The second family's two leptons are the muon and the muon neutrino; its quarks are the "charm," or *c*, quark, and the "strange," or *s*, quark. The third family's confirmed members are its two leptons—the tau lepton and the tau neutrino—and the "bottom," or *b*, quark. The remaining quark, called the "top," or *t*, quark, is crucial to the electroweak theory. The particle has not been discovered, but we and most other physicists believe it exists and presume it is simply too massive to be brought into existence by today's particle accelerators.

No members of the second and third families are stable (again, with the possible exception of the neutrinos). Their lifetimes range between a millionth and a ten-trillionth of a second, at the end of which they decay into particles of lower mass.

There are two substantial gaps in the electroweak theory's grouping of particles. First, although the theory requires that fermions come in pairs, it does not specify how many pairs constitute a family. There is no reason why each family should not have, in addition to its leptons and quarks, particles of another, still unobserved type. This possibility interests a great number of our colleagues, but so far no new particles have been observed. Second, the theory says nothing about the central question of this article: the number of families of matter. Might there be higher families made up of particles too massive for existing accelerators to produce?

At present, physicists can do nothing but insert observed masses into theories on an ad hoc basis. Some pattern can, however, be discerned [*see illustration on page 37*]. Within a given class of particle (say, a charged lepton or a quark of charge +2/3 or of −1/3), the mass increases considerably in each succeeding family.

The smallest such increase is the nearly 17-fold jump from the muon in the second family to the tau lepton in the third.

Another striking feature is found within families. Leptons are always less massive than quarks, and in every pair of leptons the neutrino is always substantially the less massive particle. In fact, it is uncertain whether neutrinos have any masses at all: experimental evidence merely puts upper limits on the mass each variety can have.

This lightness of neutrinos is essential to the method reported here for counting the number of families of particles. Even if the quark and lepton members of a fourth, fifth or sixth family were far too massive to be created by existing accelerators, the likelihood is nonetheless great that their neutrinos would have little or no mass. Almost certainly the mass of such neutrinos would be less than half the mass of the Z boson. If such neutrinos exist, therefore, they would be expected to be among the decay products of the Z, the only particle that decays copiously into pairs of neutrinos.

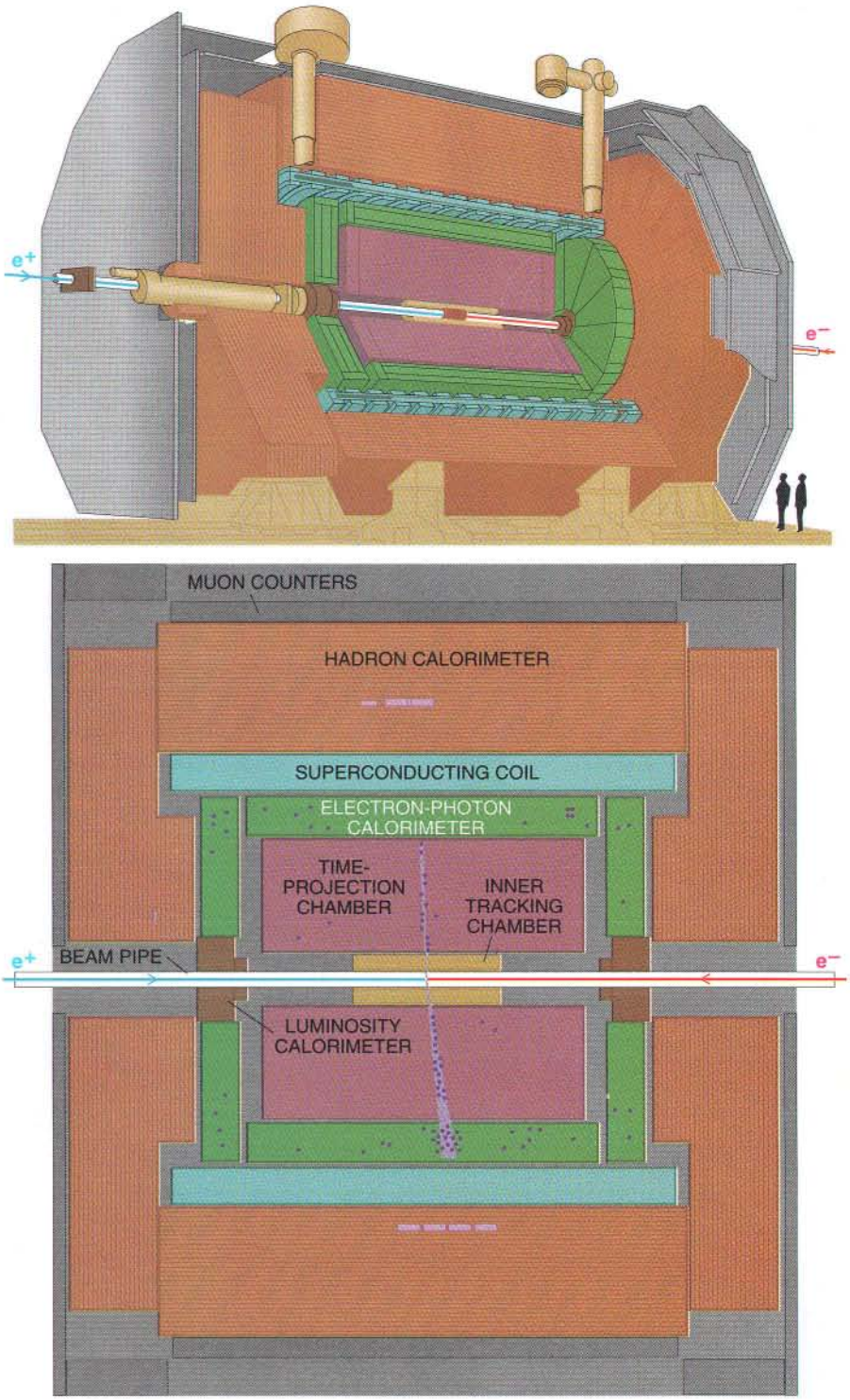
Unfortunately, neutrinos are hard to detect because they do not engage in electromagnetic or strong interactions. They touch matter only through forces that are called "weak," with good reason: most neutrinos pass through the earth without interacting. In the experiments we shall describe, the existence of neutrinos is sought indirectly.

The process begins by creating Z particles. The Z can be produced by an electron-positron pair whose combined kinetic energies make up the difference between their rest masses (expressed in equivalent energy) and the rest mass of the Z. Because these leptons have tiny rest masses, the beams in which they travel must each be raised to the very high energy of 45.5 billion electron volts (eV), about half the Z mass.

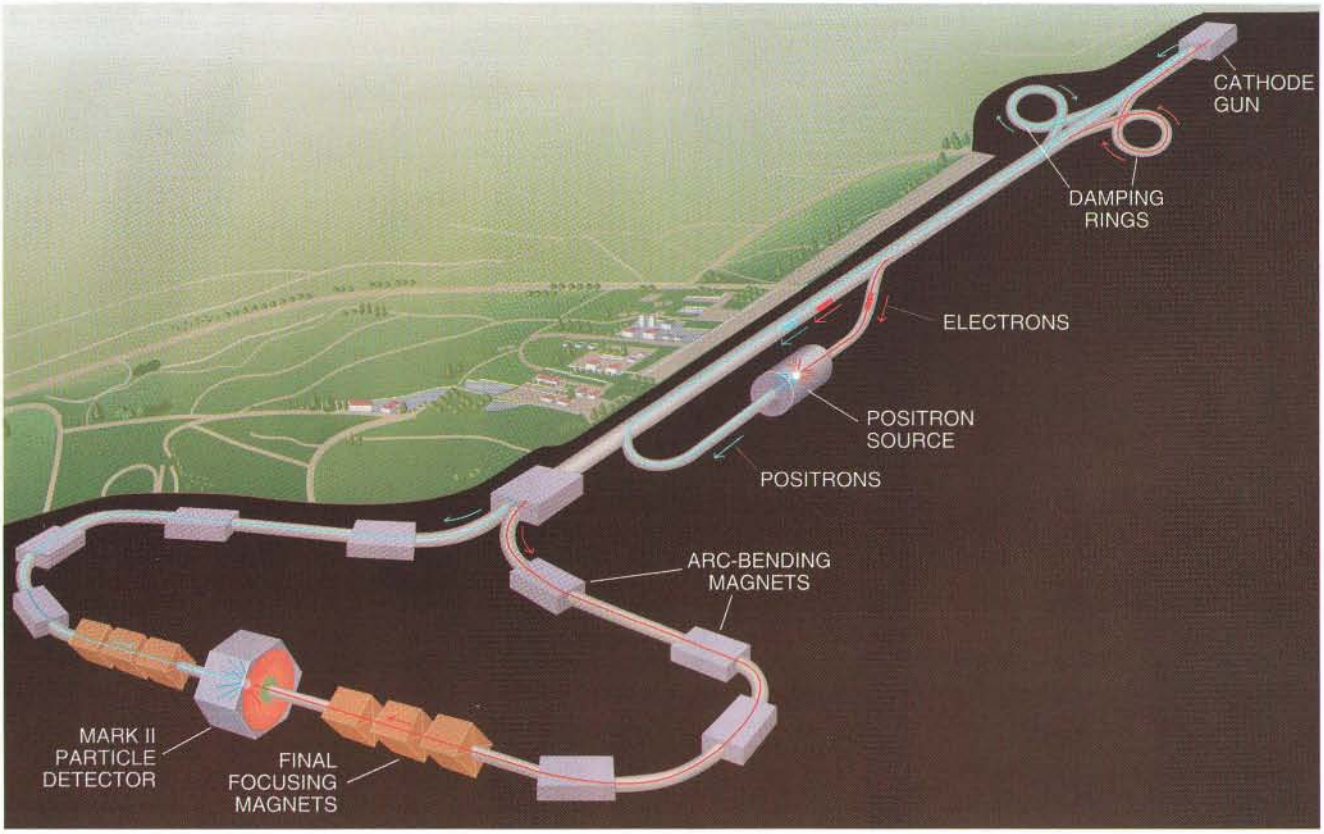
Now if the Z were perfectly stable, the beam energy would have to equal this value precisely to conserve energy and momentum. But such perfect stability is impossible, for if the Z can be created from particles, then it must also be free to decay back into them. In fact, the Z has many "channels" in which to decay. Each decay channel shortens the life of the Z.

Near the beginning of this article, we mentioned that the Z's short life made its mass indeterminate and that the extent of the indeterminacy could be used to number the families of matter. Let us explain why this must be so. One form of the Heisenberg uncertainty principle stipulates that the shorter the duration of a state is, the more uncertain its energy must be. Because the Z is short-lived, its energy—or equivalently, its mass—will have a degree of uncertainty. What this means is the following: the mass of any individual Z can be measured quite precisely, but different Zs will have slightly different masses. If the measured masses of many Zs are plotted, the resulting graph has a characteristic bell-like shape. The width of this shape is proportional to the speed at which the Z decays.

The shape is measured by varying



DETECTORS OF ALEPH are arranged in onionlike layers that feed data into computers, which can reconstruct decay events on a screen (*bottom*). Charged particles appear as tracks. The energy of both charged and neutral particles is gauged by calorimeters and graphically displayed. Aleph weighs about 4,000 tons (*top*).



STANFORD LINEAR COLLIDER speeds positrons and electrons along a three-kilometer straightaway. The injector (*top right*) shoots electrons (*red*) into a damping ring, which condenses them for later focusing. One bunch then enters the straightaway behind a bunch of positrons (*blue*). The two bunches accelerate in tandem before entering separate arcs that focus and direct them to collision in the Mark II detector (*bottom left*). Meanwhile the second bunch of electrons slams into a target, producing positrons (*center*). The positrons are returned to the front, damped and stored.

the collision energy and observing the number of Z particles produced. The measurements trace a curve that peaks, or resonates, at a combined beam energy of about 91 billion eV. This point, called the peak cross section, defines the average Z mass. The width of the resonance curve defines the particle's mass uncertainty.

The width equals the sum of partial widths contributed by each of the Z's decay channels. The known channels are the decays to particle and antiparticle pairs of all fermions with less than one half the Z mass: the three varieties of charged leptons, the five kinds of quarks and the three varieties of neutrinos. If there are other fermions whose masses are less than half the Z mass, the Z will decay to these as well, and these channels will also contribute to the Z width, making it larger.

The present experiments show that such decays to new, charged particles do not occur, so we can be sure that the particles do not exist or that their masses are larger than half the Z mass. If, however, higher-mass families do exist, then—as we argued before—their neutrinos would still be expected to have masses much smaller than half the Z mass. Therefore, the Z would also decay to these channels, and although the neutrinos would not be seen directly in these experiments, these neutrino species would contribute to the Z width and so be observable. This is the principle enabling the experiments reported here to number the families of matter.

The electroweak theory predicts the contributions of the known channels to an accuracy of about 1 percent, as follows: for the combined quark channels, 1.74 billion eV; for each charged lepton channel, 83.5 million eV; and for each neutrino channel, 166 million eV.

As the number of assumed neutrinos (and hence families) increases, the predicted Z width also increases. The predicted peak cross section, on the other hand, declines by the square of the width [*see illustration on page 39*]. One can consequently deduce the number of families either from the measured width or from the peak cross section. The latter is statistically the more powerful measurement. The establishment of the number of families by direct experimental measurement had to await the production of large numbers of Zs by the well-understood process of electron-positron annihilation.

Researchers at CERN attacked the problem by developing the Large Electron-Positron (LEP) Collider, a traditional storage-ring design built on an unprecedented scale [see "The LEP Collider," by Stephen Myers and Emilio Picasso; SCIENTIFIC AMERICAN, July 1990]. The ring, which measures 27 kilometers in circumference, is buried between 50 and 150 meters under the plain that stretches from Geneva to the French part of the Jura Mountains [*see illustration on page 34*]. Resonance cavities accelerate the two beams with radio-frequency power. The beams move in opposite directions through a roughly circular tube. Electromagnets bend the beams around every curve and direct them to collisions in four areas, each of which is provided with a large detector.

The ring design has the advantage of storing the particles indefinitely, so that they can continue to circulate and collide. It has the disadvantage of draining the beams of energy in the form of synchroton radiation, an emission

made by any charged particle that is diverted by a magnetic field. Such losses, which at these energies appear as X rays, increase as the fourth power of the beam's energy and are inversely proportional to the ring's radius. Designers can therefore increase the power of their beams by either pouring in more energy or building larger rings, or both. If optimal use is made of resources, the cost of such storage rings scales as the square of beam energy. The LEP is thought to approach the practical economic limit for accelerators of this kind.

At Stanford, the problem of making electrons and positrons collide at high energy was attacked in a novel way in the Stanford Linear Collider (SLC). The electrons and positrons are accelerated in a three-kilometer-long linear accelerator, which had been built for other purposes. They are sent into arcs a kilometer long, brought into collision and then dumped [*see illustration on opposite page*]. The electrons and positrons each lose about 2 percent of their energy because of synchrotron radiation in the arcs, but this loss is tolerable because the particles are not recirculated. A single detector is placed at the point of collision [see "The Stanford Linear Accelerator," by John R. Rees; SCIENTIFIC AMERICAN, October 1989].

The LEP is an efficient device: when the electron and positron beams recirculate, about 45,000 collisions per second occur. The SLC beams collide, at the most, only 120 times per second. Thus, the SLC must increase its efficiency. This task can be accomplished by reducing the beam's cross section to an extremely small area. The smaller the cross section of the area becomes, the more likely it is that an electron will collide head-on with a positron. The SLC has produced beam diameters of four-millionths of a meter, about one fifth the thickness of a human hair.

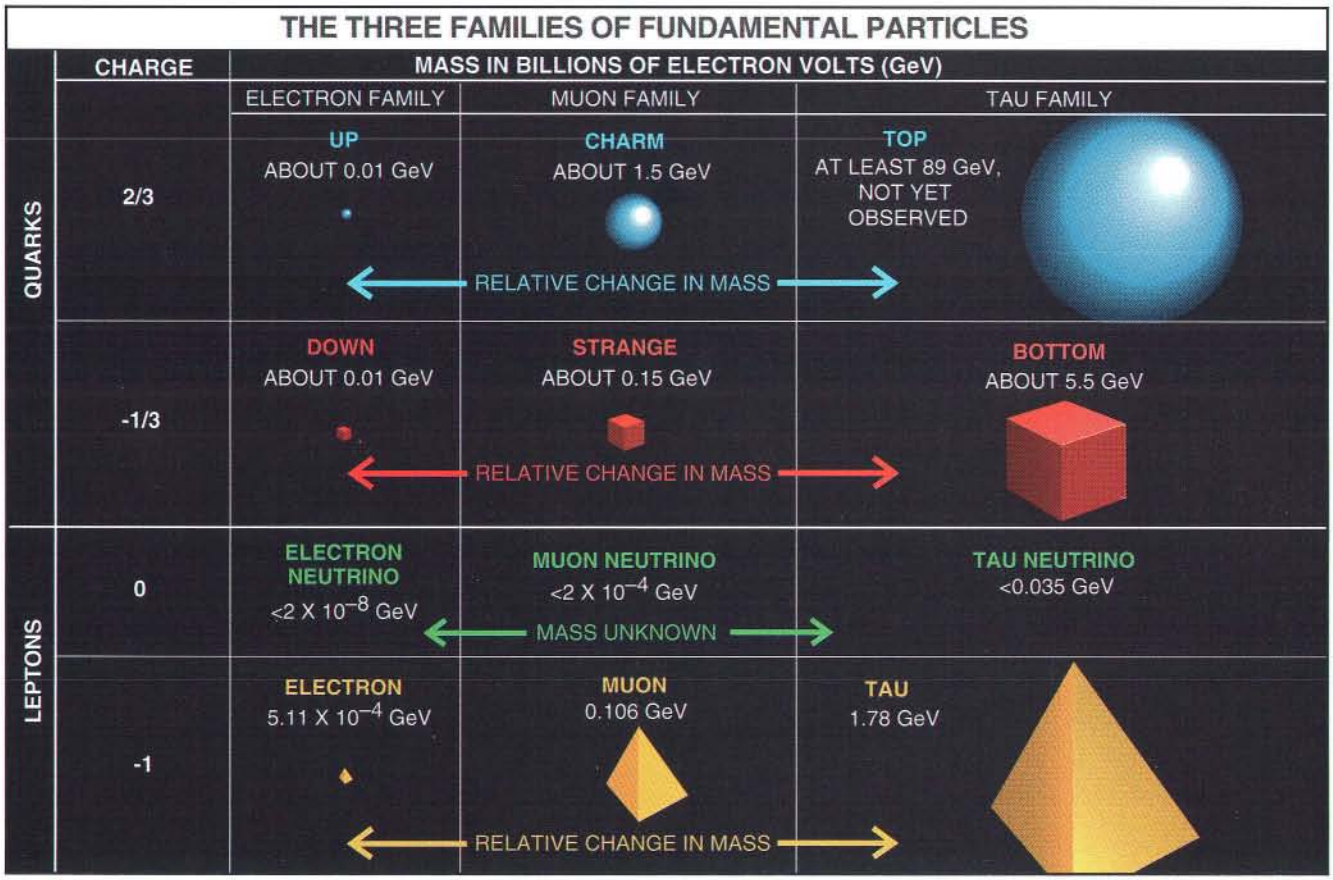
One of the main justifications for building the SLC was that it would serve as a prototype for this new kind of collider. Indeed, the SLC has shown that useful numbers of collisions are obtainable in linear colliders, and it has thus encouraged developmental research in this direction, both at SLAC and at CERN. The present Z production rates at the SLC are, however, still more than 100 times smaller than those at the LEP.

Large teams of physicists analyze the collision products in big detectors. The SLC's detector is called Mark II, and the LEP's four detectors are called Aleph, Opal, Delphi and L3 [*see illustration on page 34*]. The SLAC team numbers about 150 physicists; each of the CERN teams numbers about 400 people, drawn from research institutes and universities of two dozen countries.

The function of a detector is to measure the energies and directions of as many as possible of the particles constituting a collision event and to identify their nature, particularly that of the charged leptons. Detectors are made in onionlike layers, with tracking devices on the inside and calorimeters on the outside. Tracking devices measure the angles and momenta of charged particles. The trajectories are located by means of the ionization trails the collision products leave behind in a suitable gas. Other media, such as semiconductor detectors and light-emitting plastic fibers, are also used.

The tracking devices are generally placed in strong magnetic fields that bend the particles' trajectories inversely with respect to their momenta. Measurement of the curves yields the momenta, which in turn provide close estimates of the energy. (At the energies encountered in these experiments, the energy and the momentum of a particle differ very little.)

Calorimeters measure the energies of both neutral and charged particles by dissipating these energies in successive secondary interactions in some dense medium. This energy is then sampled in a suitable way and localized as precisely as the granularity of the calorimeter allows. Calorimeters perform their function in a number of ways. The



most common method uses sandwiches of thin sheets of dense matter, such as lead, uranium or iron, which are separated by layers of track-sensitive material.

Particles leave their mark in such materials by knocking electrons from their atoms. Argon, either in liquid form or as a gas combined with organic gases, is the usual medium. Plastic scintillators work differently: when a reaction particle traverses them, it produces a flash of light whose intensity is then measured. The calorimeter usually has two layers, an inner one optimized for the measurement of electrons and photons and an outer one optimized for hadrons.

To gather all the reaction products, the ideal detector would cover the entire solid angle surrounding the interaction point. Such detectors were pioneered in the 1970s at SLAC. In the LEP's Aleph detector the tracking of the products from the annihilation of a positron and an electron proceeds in steps.

A silicon-strip device adjoining the reaction site fixes the forward end point of each trajectory to within ten-millionths of a meter (about half the breadth of a human hair). Eight layers of detection wires then track the trajectory through an inner chamber 60 centimeters in diameter. Finally, a so-called time-projection chamber, 3.6 meters in diameter, uses a strong electric field to collect electrons knocked from gas molecules by the traversing particles. The field causes the electrons to drift to the cylindrical chambers' two ends, where they are amplified and detected on 50,000 small pads. Each electron's point of origin is inferred from the place it occupies on the pads and the time it takes to get there.

The next step outward brings the reaction products to the electron-photon calorimeter. The products traverse the superconducting coil, which creates a 15,000-gauss magnetic field at the axis of the device, and then enter the hadron calorimeter. This device, a series of iron plates separated by gas counters, also returns the magnetic flux, just as an iron core does in a conventional electromagnet. Aleph weighs 4,000 tons and cost about $60 million to build. Half a million channels of information must be read for each event, and the computer support necessary for the acquisition and later evaluation of the data is considerable [*see illustration on page 35*].

The data gathered in the first few months of operation of the two colliders have provided the best support yet adduced for the predictions of the electroweak theory. More important, they have delineated the curve describing the Z width with great precision.

The overwhelming majority of observed electron-positron annihilations give rise to four sets of products: 88 percent produce a quark and an antiquark; the remaining 12 percent are divided equally among the production of a tau lepton and antitau lepton, muon and antimuon, and electron and positron. (The last case simply reverses the initial annihilation.)

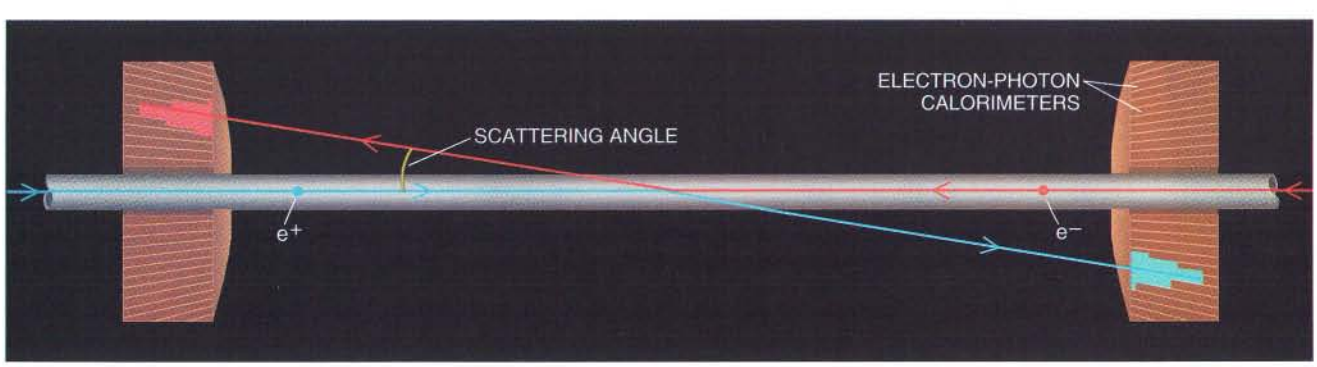
In the decays into electrons and muons, two tracks are seen back to back, with momenta (and energies) corresponding to half of the combined beam energy. The two products are easily distinguished by their distinct behavior in the calorimeters. The decays to tau leptons are more complex because they subsist for a mere instant—during which they travel about a millimeter—before decaying into tertiary particles that alone can be observed. A tau lepton leaves either closely packed tracks or just one track; in both cases, the signature is mirrored by that of another tau lepton moving in the opposite direction (thus conserving momentum).

The quarks that account for most reactions cannot be seen in their free, or "naked," state, because at birth they undergo a process called hadronization. Each quark "clothes" itself in a jet of hadrons, numbering 15 on average, two thirds of which are charged. This, the most complex of the four main decay events, usually manifests itself as back-to-back jets, each containing many tracks [*see bottom left of illustration on page 33*]. The results described here are based on the analysis of about 80,000 Z decays into quarks—the combined result of the four LEP teams and the one SLAC team.

The Z production curve is determined in an energy scan. Production probability is measured at a number of energies: at the peak energy, as well as above and below it. A precise knowledge of the beam energy is of great importance here. It was obtained at the two colliders very differently, in both cases with a good deal of ingenuity and with a precision of three parts in 10,000.

As was pointed out earlier, the total width of the Z resonance can be determined from either the height at the peak energy or the width of the resonance curve. The height has the smaller statistical error but requires knowledge not only of the rate at which events occur but also of the rate at which particles from the two beams cross. The latter rate is called the luminosity of the collider.

In the simple case of two perfectly aligned beams of identical shape and size, the luminosity equals the product of the number of electrons and the number of positrons in each crossing bunch, multiplied by the number of bunches crossing each second, divid-



ALEPH'S LUMINOSITY DETECTOR registers a small-angle scattering event when a positron (e^+) enters from the left and glances off an electron (e^-) entering from the right. The particles then hurtle into fine-grained calorimeters that fix their angles and measure their energies. The rate of these events measures the LEP's collision frequency, or luminosity. One must know the luminosity to determine how changes in beam energy affect the probability of producing Z bosons. This probability function, in turn, predicts the number of neutrino varieties, hence the number of families of matter.

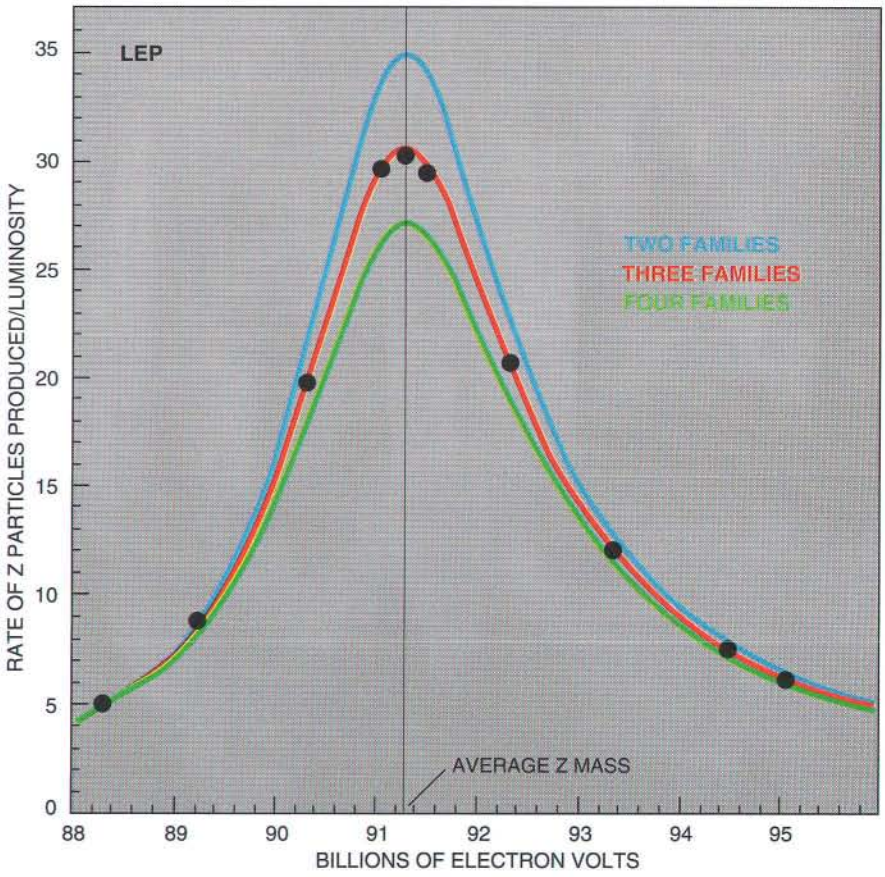
ed by the cross-sectional area of the beams. In practice, luminosity is determined only by observing the rate of the one process that is known with precision: the scattering of electrons and positrons that glance off one another at very small angles without combining or otherwise changing state. To record such so-called elastic collisions, two special detectors are placed in small angular regions just off the axis of the beam pipe. One of the detectors is in front of the collision area; the other is behind it. In the case of Aleph, these detectors are electron-photon calorimeters of high granularity [*see illustration on opposite page*].

The elastically scattered electrons and positrons are identified by the characteristic pattern in which they deposit energy in the detectors and by the way they strike the two detectors back to back, producing a perfectly aligned path. The essence here is to understand precisely the way in which particles are registered, especially in those parts of the detectors that correspond to exceedingly small scattering angles. This is important because the detection rate is extremely sensitive to changes in the angle.

When the resulting data are fitted to the theoretical resonance shape, three parameters are considered: the height at the peak, the total width and the *Z* mass. The data, in fact, agree well with the shape of the theoretically expected distribution. The next step, then, is to determine the number of neutrino families from two independent parameters—the width and the peak height.

The combined results of the five teams produced an average estimate of 3.09 neutrino varieties, with an experimental uncertainty of 0.09. This number closely approaches an integer, as it should, and matches the number of neutrino varieties that are already known. A fourth neutrino could exist without contradicting these findings only if its mass exceeded 40 billion eV—a most unlikely possibility, given the immeasurably small masses of the three known neutrinos.

The *Z* result fits the cosmological evidence gathered by those who study matter on galactic and supergalactic scales. Astronomers have measured the ratio of hydrogen to helium and other light elements in the universe. Cosmologists and astrophysicists have tried to infer the processes by which these relative abundances came about [see "Particle Accelerators Test Cosmological Theory," by David N. Schramm and Gary Steigman; SCIENTIFIC AMERICAN, June 1988].



RESONANCE CURVES predicted for the *Z* particle vary according to the number of families of matter. Thousands of *Z* decays into quarks, observed at CERN, appear as points. The measurements agree with the expectation for three families of matter.

Shortly after the big bang, the cataclysmic explosion that created the universe and began its expansion, matter was so hot that a neutron was as likely to decay into a proton-electron pair as the latter was to combine to form a neutron. Consequently, as many neutrons as protons existed. But as the universe expanded and cooled, the slightly heavier neutrons changed into protons more readily than protons changed into neutrons. The neutron-proton ratio therefore fell steadily.

When the expansion brought the temperature of the universe below one billion kelvins, protons and neutrons were for the first time able to fuse, thereby forming some of the lighter elements, mainly helium. The resulting abundances depend critically on the ratio of neutrons to protons at the time light elements were forming. This ratio, in turn, depends on the rate at which the universe expanded and cooled. At this stage, each light neutrino family—that is, any whose constituents have a mass smaller than about a million eV—contributes appreciably to the energy density and cooling rate. The measured abundances of light elements are consistent with cosmological models that assume the existence of three light neutrino families but tend to disfavor those that assume four or more.

Many questions remain unanswered. Why are there just three families of particles? What law determines the masses of their members, decreeing that they shall span 10 powers of 10? These problems lie at the center of particle physics today. They have been brought one step closer to solution by the numbering of the families of matter.

FURTHER READING

QUARKS WITH COLOR AND FLAVOR. Sheldon Lee Glashow in *Scientific American*, Vol. 233, No. 4, pages 38–50; October 1975.

QUARKS: THE STUFF OF MATTER. Harald Fritzsch. Basic Books, 1983.

ELEMENTARY PARTICLES AND FORCES. Chris Quigg in *Scientific American*, Vol. 252, No. 4, pages 84–95; April 1985.

THE COSMIC ONION. Frank E. Close. American Institute of Physics, 1988.

THE EXPERIMENTAL FOUNDATIONS OF PARTICLE PHYSICS. Robert N. Cahn and Gerson Goldhaber. Cambridge University Press, 1989.

On the Generalized Theory of Gravitation

An account of the newly published extension of the general theory of relativity against its historical and philosophical background

by Albert Einstein

The editors of SCIENTIFIC AMERICAN have asked me to write about my recent work which has just been published. It is a mathematical investigation concerning the foundations of field physics.

Some readers may be puzzled: Didn't we learn all about the foundations of physics when we were still at school? The answer is "yes" or "no," depending on the interpretation. We have become acquainted with concepts and general relations that enable us to comprehend an immense range of experiences and make them accessible to mathematical treatment. In a certain sense these concepts and relations are probably even final. This situation is true, for example, of the laws of light refraction, of the relations of classical thermodynamics as far as it is based on the concepts of pressure, volume, temperature, heat and work, and of the hypothesis of the nonexistence of a perpetual motion machine.

What, then, impels us to devise theory after theory? Why do we devise theories at all? The answer to the latter question is simply because we enjoy "comprehending," that is, reducing phenomena by the process of logic to something already known or (apparently) evident. New theories are first of all necessary when we encounter new facts that cannot be "explained" by existing theories. But this motivation for setting up new theories is, so to speak, trivial, imposed from without. There is another, more subtle motive of no less importance. This is the striving toward unification and simplification of the premises of the theory as a whole (that is, Mach's principle of economy, interpreted as a logical principle).

There exists a passion for comprehension, just as there exists a passion for music. That passion is rather common in children but gets lost in most people later on. Without this passion, there would be neither mathematics nor natural science. Time and again the passion for understanding has led to the illusion that man is able to comprehend the objective world rationally, by pure thought, without any empirical foundations—in short, by metaphysics. I believe that every true theorist is a kind of tamed metaphysicist, no matter how pure a "positivist" he may fancy himself. The metaphysicist believes that the logically simple is also the real. The tamed metaphysicist believes that not all that is logically simple is embodied in experienced reality but that the totality of all sensory experience can be "comprehended" on the basis of a conceptual system built on premises of great simplicity. The skeptic will say that this is a "miracle creed." Admittedly so, but it is a miracle creed which has been borne out to an amazing extent by the development of science.

The rise of atomism is a good example. How may Leucippus have conceived this bold idea? When water freezes and becomes ice—apparently something entirely different from water—why is it that the thawing of the ice forms something that seems indistinguishable from the original water? Leucippus is puzzled and looks for an "explanation." He is driven to the conclusion that in these transitions the "essence" of the thing has not changed at all. Maybe the thing consists of immutable particles, and the change is only a change in their spatial arrangement. Could it not be that the same is true of all material objects which emerge again and again with nearly identical qualities?

This idea is not entirely lost during the long hibernation of occidental thought. Two thousand years after Leucippus, Bernoulli wonders why gas exerts pressure on the walls of a container. Should this phenomenon be "explained" by mutual repulsion of the parts of the gas, in the sense of Newtonian mechanics? This hypothesis appears absurd, for the gas pressure depends on the temperature, all other things being equal. To assume that the Newtonian forces of interaction depend on temperature is contrary to the spirit of Newtonian mechanics. Since Bernoulli is aware of the concept of atomism, he is bound to conclude that the atoms (or molecules) collide with the walls of the container and in doing so exert pressure. After all, one has to assume that atoms are in motion; how else can one account for the varying temperature of gases?

A simple mechanical consideration shows that this pressure depends only on the kinetic energy of the particles and on their density in space. This should have led the physicists of that age to the conclusion that heat consists in random motion of the atoms. Had they taken this consideration as seriously as it deserved to be taken, the development of the theory of heat—in particular the discovery of the

ALBERT EINSTEIN (1879–1955) formulated the general theory of relativity and the special theory of relativity and profoundly influenced modern physics in many other ways. He was awarded the Nobel Prize for Physics in 1921.

Reprinted from the April 1950 issue

equivalence of heat and mechanical energy—would have been considerably facilitated.

This example is meant to illustrate two things. The theoretical idea (atomism in this case) does not arise apart from and independent of experience; nor can it be derived from experience by a purely logical procedure. It is produced by a creative act. Once a theoretical idea has been acquired, one does well to hold fast to it until it leads to an untenable conclusion.

As for my latest theoretical work, I do not feel justified in giving a detailed account of it before a wide group of readers interested in science. That should be done only with theories which have been adequately confirmed by experience. So far it is primarily the simplicity of its premises and its intimate connection with what is already known (namely, the laws of the pure gravitational field) that speak in favor of the theory to be discussed here. It may, however, be of interest to a wide group of readers interested in science to become acquainted with the train of thought that can lead to endeavors of such an extremely speculative nature. Moreover, it will be shown what kinds of difficulties are encountered and in what sense they have been overcome.

In Newtonian physics the elementary theoretical concept on which the theoretical description of material bodies is based is the material point, or particle. Thus, matter is considered a priori to be discontinuous. This assumption makes it necessary to consider the action of material points on one another as "action at a distance." Because the latter concept seems quite contrary to everyday experience, it is only natural that the contemporaries of Newton—and indeed Newton himself—found it difficult to accept. Because of the almost miraculous success of the Newtonian system, however, the succeeding generations of physicists became accustomed to the idea of action at a distance. Any doubt was buried for a long time to come.

But when, in the second half of the 19th century, the laws of electrodynamics became known, it turned out that these laws could not be satisfactorily incorporated into the Newtonian system. It is fascinating to muse: Would Faraday have discovered the law of electromagnetic induction if he had received a regular college education? Unencumbered by the traditional way of thinking, he felt that the introduction of the "field" as an independent element of reality helped him to coordinate the experimental facts. It was Maxwell who fully comprehended the significance of the field concept; he made the fundamental discovery that the laws of electrodynamics found their natural expression in the differential equations for the electric and magnetic fields. These equations implied the existence of waves, whose properties corresponded to those of light as far as they were known at that time. This incorporation of optics into the theory of electromagnetism represents one of the greatest triumphs in the striving toward unification of the foundations of physics. Maxwell achieved this unification by purely theoretical arguments, long before it was corroborated by Hertz's experimental work. The new insight made it possible to dispense with the hypothesis of action at a distance, at least in the realm of electromagnetic phenomena; the intermediary field now appeared as the only carrier of electromagnetic interaction between bodies, and the field's behavior was completely determined by contiguous processes, expressed by differential equations.

Now a question arose: Since the field exists even in a vacuum, should one conceive of the field as a state of a "carrier," or should it rather be endowed with an independent existence not reducible to anything else? In other words, is there an "ether" which carries the field; the ether being considered in the undulatory state, for example, when it carries light waves?

The question has a natural answer: because one cannot dispense with the field concept, it is preferable not to introduce in addition a carrier with hypothetical properties. However, the pathfinders who first recognized the indispensability of the field concept were still too strongly imbued with the mechanistic tradition of thought to accept unhesitatingly this simple point of view. But during the course of the following decades this view imperceptibly took hold.

The introduction of the field as an elementary concept gave rise to an inconsistency of the theory as a whole. Maxwell's theory, although adequately describing the behavior of electrically charged particles in their interaction with one another, does not explain the behavior of electrical densities, that is, it does not provide a theory of the particles themselves. They must therefore be treated as mass points on the basis of the old theory. The combination of the idea of a continuous field with the conception of material points discontinuous in space appears inconsistent. A consistent field theory requires continuity of all elements of the theory, not only in time but also in space and in all points of space. Hence, the material particle has no place as a fundamental concept in a field theory. Thus, even apart from the fact that gravitation is not included, Maxwell's electrodynamics cannot be considered a complete theory.

Maxwell's equations for empty space remain unchanged if the spatial coordinates and the time are subjected to a particular kind of linear transformations—the Lorentz transformations ("covariance" with respect to Lorentz transformations). Covariance also holds, of course, for a transformation composed of two or more such transformations; this is called the "group" property of Lorentz transformations.

Maxwell's equations imply the "Lorentz group," but the Lorentz group does not imply Maxwell's equations. Indeed, it is possible to redefine the Lorentz group independently of Maxwell's equations as a group of linear transformations that leave a particular value of the velocity—the velocity of light—invariant. These transformations hold for the transition from one "inertial system" to another which is in uniform motion relative to the first. The most conspicuous novel property of this transformation group is that it does away with the absolute character of the concept of simultaneity of events distant from one another in space. On this account it is to be expected that all equations of physics are covariant with respect to Lorentz transformations (special theory of relativity). Thus it came about that Maxwell's equations led to a heuristic principle valid far beyond the range of the applicability or even validity of the equations themselves.

Special relativity has this in common with Newtonian mechanics: the laws of both theories are supposed to hold only with respect to certain coordinate systems—those known as inertial systems. An inertial system is a system in a state of motion such that "force-free" material points within it are not accelerated with respect to the coordinate system. Yet this defini-

tion is empty if there is no independent means for recognizing the absence of forces. But such a means of recognition does not exist if gravitation is considered as a "field."

Let A be a system uniformly accelerated with respect to an inertial system I. Material points, not accelerated with respect to I, are accelerated with respect to A, the acceleration of all the points being equal in magnitude and direction. They behave as if a gravitational field exists with respect to A, for it is a characteristic property of the gravitational field that the acceleration is independent of the particular nature of the body.

There is no reason to exclude the possibility of interpreting this behavior as the effect of a "true" gravitational field (*principle of equivalence*). This interpretation implies that A is an inertial system, even though it is accelerated with respect to another inertial system. (It is essential for this argument that the introduction of independent gravitational fields is considered justified even though no masses generating the field are defined. Therefore, to Newton such an argument would not have appeared convincing.)

Thus, the concepts of inertial system, the law of inertia and the law of motion are deprived of their concrete meaning—not only in classical mechanics but also in special relativity. Moreover, following up this train of thought, it turns out that with respect to A time cannot be measured by identical clocks; indeed, even the immediate physical significance of coordinate differences is generally lost. In view of all these difficulties, should one not try, after all, to hold on to the concept of the inertial system, relinquishing the attempt to explain the fundamental character of the gravitational phenomena that manifest themselves in the Newtonian system as the equivalence of inert and gravitational mass? Those who trust in the comprehensibility of nature must answer: No.

This is the gist of the principle of equivalence: in order to account for the equality of inert and gravitational mass within the theory, it is necessary to admit nonlinear transformations of the four coordinates. That is, the group of Lorentz transformations, and hence the set of the "permissible" coordinate systems, has to be extended.

What group of coordinate transformations can then be substituted for the group of Lorentz transformations? Mathematics suggests an answer which is based on the fundamental investigations of Gauss and Riemann: namely, that the appropriate substitute is the group of all continuous (analytical) transformations of the coordinates. Under these transformations the only thing that remains invariant is the fact that neighboring points have nearly the same coordinates; the coordinate system expresses only the topological order of the points in space (including its four-dimensional character). The equations expressing the laws of nature must be covariant with respect to all continuous transformations of the coordinates. This is the principle of general relativity.

The procedure just described overcomes a deficiency in the foundations of mechanics that had already been noticed by Newton and was criticized by Leibnitz and, two centuries later, by Mach: inertia resists acceleration, but acceleration relative to what? Within the frame of classical mechanics the only answer is: inertia resists acceleration relative to space. This is a physical property of space—space acts on objects, but objects do not act on space. Such is probably the deeper meaning of Newton's assertion *spatium est absolutum* (space is absolute). But the idea disturbed some, in particular Leibnitz, who did not ascribe an independent existence to space but considered it merely a property of "things" (contiguity of physical objects). Had his justified doubts won out at that time, it hardly would have been a boon to physics, inasmuch as the empirical and theoretical foundations necessary to follow up his idea were not available in the 17th century.

According to general relativity, the concept of space detached from any physical content does not exist. The physical reality of space is represented by a field whose components are continuous functions of four independent variables—the coordinates of space and time. It is just this particular kind of dependence that expresses the spatial character of physical reality.

Since the theory of general relativity implies the representation of physical reality by a *continuous* field, the concept of particles or material points cannot play a fundamental part, and neither can the concept of motion. The particle can only appear as a limited region in space in which the field strength or the energy density is particularly high.

A relativistic theory has to answer two questions: namely, What is the mathematical character of the field? and What equations hold for this field?

Concerning the first question: From the mathematical point of view, the field is essentially characterized by the way its components transform if a coordinate transformation is applied. Concerning the second question: The equations must determine the field *to a sufficient extent* while satisfying the postulates of general relativity. Whether or not this requirement can be satisfied depends on the choice of the field type.

The attempt to comprehend the correlations among the empirical data on the basis of such a highly abstract program may at first appear almost hopeless. The procedure amounts, in fact, to putting the question: What most simple property can be required from what most simple object (field) while preserving the principle of general relativity? Viewed from the standpoint of formal logic, the dual character of the question appears calamitous, quite apart from the vagueness of the concept "simple." Moreover, from the standpoint of physics, there is nothing to warrant the assumption that a theory that is "logically simple" should also be "true."

Yet every theory is speculative. When the basic concepts of a theory are comparatively "close to experience" (for example, the concepts of force, pressure, mass), its speculative character is not so easily discernible. If, however, a theory is such as to require the application of complicated logical processes in order to reach conclusions from the premises that can be confronted with observation, everybody becomes conscious of the speculative nature of the theory. In such a case an almost irresistible feeling of aversion arises in people who are inexperienced in epistemological analysis and who are unaware of the precarious nature of theoretical thinking in those fields with which they are familiar.

On the other hand, it must be conceded that a theory has an important advantage if its basic concepts and fundamental hypotheses are "close to experience," and greater confidence in such a theory is certainly justified. There is less danger of going completely astray, particularly since it takes so much less time and effort to disprove such theories by experience. Yet more and more, as the depth of our knowledge increases, we must give up this advantage in our quest for logical simplicity and uniformity in the foundations of physical theory. It has to be admitted that general relativity has gone further than previous physical theories


Ben Shahn

in relinquishing "closeness to experience" of fundamental concepts in order to attain logical simplicity. This holds already for the theory of gravitation, and it is even more true of the new generalization, which is an attempt to comprise the properties of the total field. In the generalized theory the procedure of deriving from the premises of the theory conclusions that can be confronted with empirical data is so difficult that so far no such result has been obtained. In favor of this theory are, at this point, its logical simplicity and its "rigidity." Rigidity means here that the theory is either true or false, but not modifiable.

The greatest inner difficulty impeding the development of the theory of relativity is the dual nature of the problem, indicated by the two questions we have asked. This duality is the reason why the development of the theory has taken place in two steps so widely separated in time. The first of these steps, the theory of gravitation, is based on the principle of equivalence discussed above and rests on the following consideration: According to the theory of special relativity, light has a constant velocity of propagation. If a light ray in a vacuum starts from a point, designated by the coordinates x_1, x_2 and x_3 in a three-dimensional coordinate system, at the time x_4, it spreads as a spherical wave and reaches a neighboring point $(x_1 + dx_1, x_2 + dx_2, x_3 + dx_3)$ at the time $x_4 + dx_4$. Introducing the velocity of light, c, we write the expression:

$$\sqrt{dx_1^2 + dx_2^2 + dx_3^2} = c dx_4$$

This can also be written in the form:

$$dx_1^2 + dx_2^2 + dx_3^2 - c^2 dx_4^2 = 0$$

This expression represents an objective relation between neighboring space-time points in four dimensions, and it holds for all inertial systems, provided the coordinate transformations are restricted to those of special relativity. The relation loses this form, however, if arbitrary continuous transformations of the coordinates are admitted in accordance with the principle of general relativity. The relation then assumes the more general form:

$$\sum_{ik} g_{ik}\, dx_i\, dx_k = 0$$

The g_{ik} are certain functions of the coordinates which transform in a definite way if a continuous coordinate transformation is applied. According to the principle of equivalence, these g_{ik} functions describe a particular kind of gravitational field: a field that can be obtained by transformation of "field-free" space. The g_{ik} satisfy a particular law of transformation. Mathematically speaking, they are the components of a "tensor" with a property of symmetry which is preserved in all transformations; the symmetrical property is expressed as follows:

$$g_{ik} = g_{ki}$$

The idea suggests itself: May we not ascribe objective meaning to such a symmetrical tensor, even though the field *cannot* be obtained from the empty space of special relativity by a mere coordinate transformation? Although we cannot expect that such a symmetrical tensor will describe the most general field, it may well describe the particular case of the "pure gravitational field." Thus it is evident what kind of field, at least for a special case, general relativity has to postulate: a symmetrical tensor field.

Hence, only the second question is left: What kind of general covariant field law can be postulated for a symmetrical tensor field?

This question has not been difficult to answer in our time, since the necessary mathematical conceptions were already at hand in the form of the metric theory of surfaces, created a century ago by Gauss and extended by Riemann to manifolds of an arbitrary number of dimensions. The result of this purely formal investigation has been amazing in many respects. The differential equations that can be postulated as field law for g_{ik} cannot be of lower than second order, that is, they must at least contain the second derivatives of the g_{ik} with respect to the coordinates. Assuming that no higher than second derivatives appear in the field law, *it is mathematically determined by the principle of general relativity.* The system of equations can be written in the form:

$$R_{ik} = 0$$

The R_{ik} transform in the same manner as the g_{ik}, that is, they too form a symmetrical tensor.

These differential equations completely replace the Newtonian theory of the motion of celestial bodies, provided the masses are represented as singularities of the field. In other words, they contain the law of force as well as the law of motion while eliminating inertial systems.

The fact that the masses appear as singularities indicates that these masses themselves cannot be explained by symmetrical g_{ik} fields, or "gravitational fields." Not even the fact that only *positive* gravitating masses exist can be deduced from this theory. Evidently a complete relativistic field theory must be based on a field of more complex nature, that is, a generalization of the symmetrical tensor field.

Before considering such a generalization, two remarks pertaining to gravitational theory are essential for the explanation to follow.

The first observation is that the principle of general relativity imposes exceedingly strong restrictions on the theoretical possibilities. Without this restrictive principle it would be practically impossible for anybody to hit on the gravitational equations, not even by using the principle of special relativity, even though one knows that the field has to be described by a symmetrical tensor. No amount of collection of facts could lead to these equations unless the principle of general relativity were used.

This is the reason why all attempts to obtain a deeper knowledge of the foundations of physics seem doomed to me unless the basic concepts are in accordance with general relativity from the beginning. This situation makes it difficult to use our empirical knowledge, however comprehensive, in looking for the fundamental concepts and relations of physics, and it forces us to apply free speculation to a much greater extent than is currently assumed by most physicists.

I do not see any reason to assume that the heuristic significance of the principle of general relativity is restricted to gravitation and that the rest of physics can be dealt with separately on the basis of special relativity, with the hope that later on the whole may be fitted consistently into a general relativistic scheme. I do not think that such an attitude, although historically understandable, can be objectively justified. The comparative smallness of what we know today as gravitational effects is not a conclusive reason for ignoring the principle of general relativity in theoretical investigations of a fundamental character. In other words, I do not believe that it is justifiable to

ask: What would physics look like without gravitation?

The second point we must note is that the equations of gravitation are 10 differential equations for the 10 components of the symmetrical tensor g_{ik}. In the case of a nongeneral relativistic theory, a system is ordinarily not overdetermined if the number of equations is equal to the number of unknown functions. The manifold of solutions is such that within the general solution a certain number of functions of three variables can be chosen arbitrarily. For a general relativistic theory, this cannot be expected as a matter of course. Free choice with respect to the coordinate system implies that out of the 10 functions of a solution, or components of the field, four can be made to assume prescribed values by a suitable choice of the coordinate system. In other words, the principle of general relativity implies that the number of functions to be determined by differential equations is not 10 but $10 - 4 = 6$. For these six functions only six independent differential equations may be postulated. Only six out of the 10 differential equations of the gravitational field ought to be independent of one another, while the remaining four must be connected to those six by means of four relations (identities). And indeed there exist among the left-hand sides, R_{ik}, of the 10 gravitational equations four identities—known as "Bianchi's identities"—which assure their "compatibility."

In a case like this—when the number of field variables is equal to the number of differential equations—compatibility is always assured if the equations can be obtained from a variational principle. This is indeed the case for the gravitational equations.

However, the 10 differential equations cannot be entirely replaced by six. The system of equations is indeed "overdetermined," but because of the existence of the identities it is overdetermined in such a way that its compatibility is not lost, that is, the manifold of solutions is not critically restricted. The fact that the equations of gravitation imply the law of motion for the masses is intimately connected with this (permissible) overdetermination.

After this preparation it is now easy to understand the nature of the present investigation without entering into the details of its mathematics. The problem is to set up a relativistic theory for the total field. The most important clue to its solution is that there exists already the solution for the special case of the pure gravitational field. The theory we are looking for must therefore be a generalization of the theory of the gravitational field. The first question is: What is the natural generalization of the symmetrical tensor field?

This question cannot be answered by itself, but only in connection with the other question: What generalization of the field is going to provide the most natural theoretical system? The answer on which the theory under discussion is based is that the symmetrical tensor field must be replaced by a nonsymmetrical one. This change means that the condition $g_{ik} = g_{ki}$ for the field components must be dropped. In that case the field has 16 instead of 10 independent components.

There remains the task of setting up the relativistic differential equations for a nonsymmetrical tensor field. In the attempt to solve this problem, one meets with a difficulty that does not arise in the case of the symmetrical field. The principle of general relativity does not suffice to determine completely the field equations, mainly because the transformation law of the symmetrical part of the field alone does not involve the components of the antisymmetrical part or vice versa. Probably this is the reason why this kind of generalization of the field has hardly ever been tried before. The combination of the two parts of the field can only be shown to be a natural procedure if in the formalism of the theory only the total field plays a role, and not the symmetrical and antisymmetrical parts separately.

It turned out that this requirement can indeed be satisfied in a natural way. But even this requirement, together with the principle of general relativity, is still not sufficient to determine uniquely the field equations. Let us remember that the system of equations must satisfy a further condition: the equations must be compatible. It has been mentioned above that this condition is satisfied if the equations can be derived from a variational principle.

This has indeed been achieved, although not in so natural a way as in the case of the symmetrical field. It has been disturbing to find that it can be achieved in two different ways. These variational principles furnished two systems of equations—let us denote them by E_1 and E_2—which were different from each other (although only slightly so), each of them exhibiting specific imperfections. Consequently, even the condition of compatibility was insufficient to determine the system of equations uniquely.

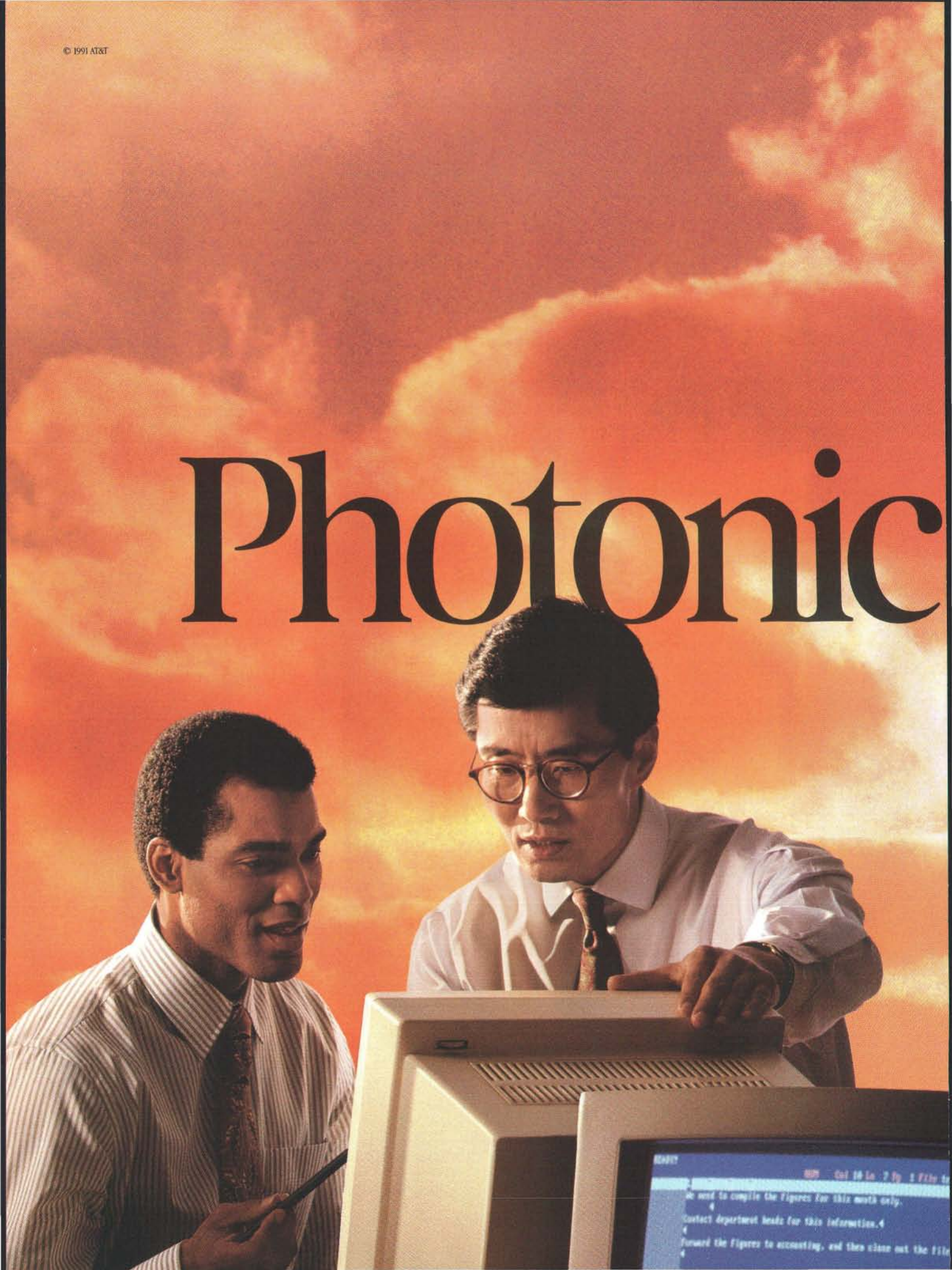
It was, in fact, the formal defects of the systems E_1 and E_2 that indicated a possible way out. There exists a third system of equations, E_3, which is free of the formal defects of the systems E_1 and E_2 and represents a combination of them in the sense that every solution of E_3 is a solution of E_1 as well as of E_2. This suggests that E_3 may be the system we have been looking for. Why not postulate E_3, then, as the system of equations? Such a procedure is not justified without further analysis, since the compatibility of E_1 and that of E_2 do not imply compatibility of the stronger system E_3, where the number of equations exceeds the number of field components by four.

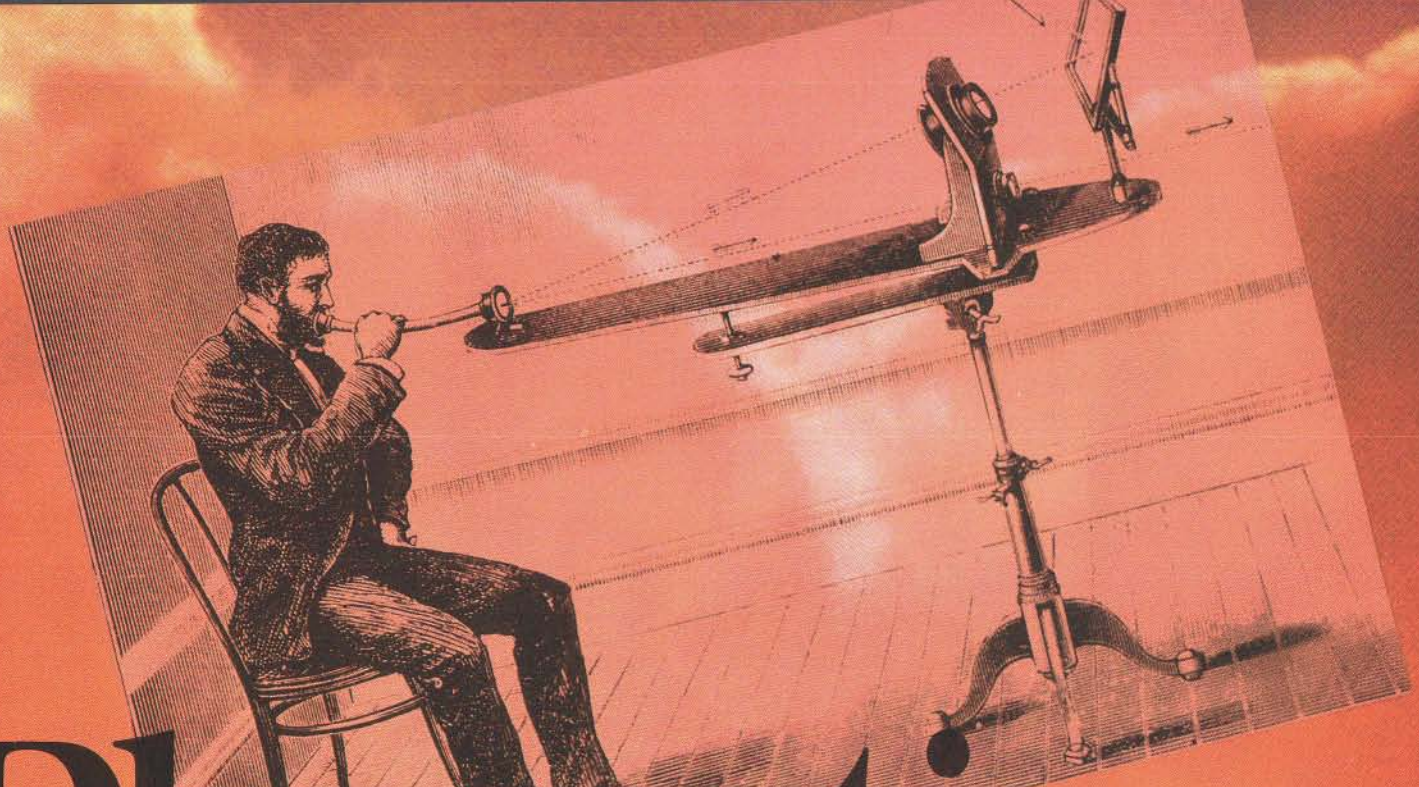
An independent consideration shows that irrespective of the question of compatibility the stronger system, E_3, is the only really natural generalization of the equations of gravitation.

But E_3 is not a compatible system in the same sense as are the systems E_1 and E_2, whose compatibility is assured by a sufficient number of identities, which means that every field that satisfies the equations for a definite value of the time has a continuous extension representing a solution in four-dimensional space. The system E_3, however, is not extensible in the same way. Using the language of classical mechanics, we might say of this situation: In the case of the system E_3 the "initial condition" cannot be freely chosen. What really matters is the answer to the question: Is the manifold of solutions for the system E_3 as extensive as must be required for a physical theory? This purely mathematical problem is as yet unsolved.

The skeptic will say: "It may well be true that this system of equations is reasonable from a logical standpoint. But this does not prove that it corresponds to nature." You are right, dear skeptic. Experience alone can decide on truth. Yet we have achieved something if we have succeeded in formulating a meaningful and precise question. Affirmation or refutation will not be easy, in spite of an abundance of known empirical facts. The derivation, from the equations, of conclusions which can be confronted with experience will require painstaking efforts and probably new mathematical methods.

Photonic





Phrontiers

Or, How Bell Labs Takes Fiber Optics Where It's Never Been Before.

When Alexander Graham Bell first discovered photonics over 100 years ago, who knew where it would lead? AT&T advances in fiber optics have made communications faster. Clearer. Now AT&T is developing the next milestone in photonics: photonic switching. A new technology to switch calls from fiber to fiber using beams of light. Photonic switches will eliminate potential bottlenecks. Allow the central office to switch terabits of bandwidth instead of gigabits. With the combination of fiber optics and photonic switching, you'll be able to receive up to 20,000 studio quality TV channels over the same lines that connect your phones. Watch your networked computers fly at unprecedented speeds. For even greater productivity. Greater savings. Learn more about what this new photonic frontier means to you. Call AT&T Network Systems at 1 800 638-7978, ext. 1210.

AT&T Network Systems and Bell Laboratories
Technologies For The Real World.

AT&T
Network Systems

The Inflationary Universe

A new theory of cosmology suggests that the observable universe is embedded in a much larger region of space that had an extraordinary growth spurt a fraction of a second after the primordial big bang

by Alan H. Guth and Paul J. Steinhardt

In the past few years certain flaws in the standard big bang theory of cosmology have led to the development of a new model of the very early history of the universe. The model, known as the inflationary universe, agrees precisely with the generally accepted description of the observed universe for all times after the first 10^{-30} second. For this first fraction of a second, however, the story is dramatically different. According to the inflationary model, the universe had a brief period of extraordinarily rapid inflation, or expansion, during which its diameter increased by a factor perhaps 10^{50} times larger than had been thought. In the course of this stupendous growth spurt, all the matter and energy in the universe could have been created from virtually nothing. The inflationary process also has important implications for the present universe. If the new model is correct, the observed universe is only a very small fraction of the entire universe.

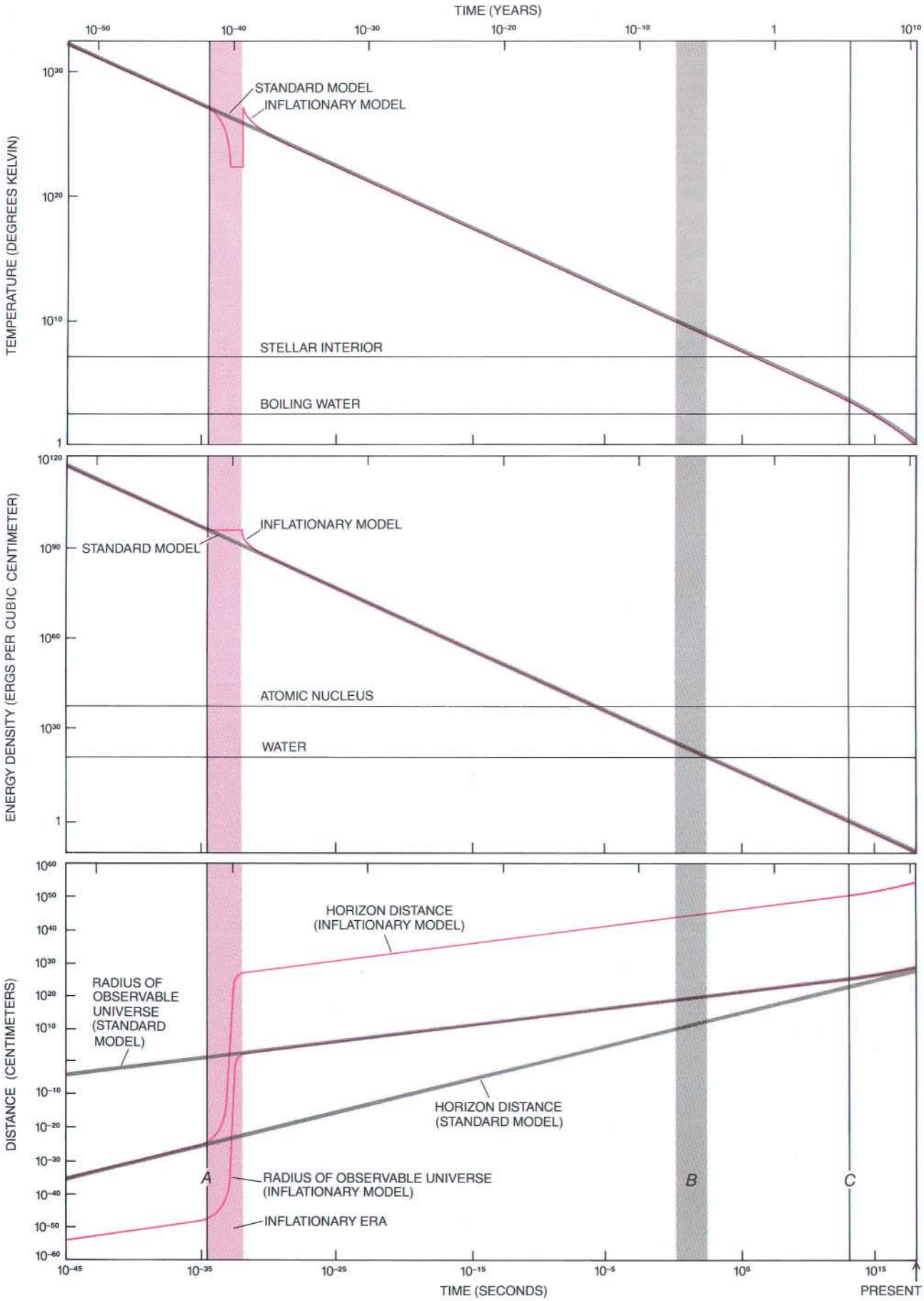
The inflationary model has many features in common with the standard big bang model. In both models the universe began between 10 and 15 billion years ago as a primeval fireball of extreme density and temperature, and it has been expanding and cooling ever since. This picture has been successful in explaining many aspects of the observed universe, including the redshifting of the light from distant galaxies, the cosmic microwave background radiation and the primordial abundances of the lightest elements. All these predictions have to do only with events that presumably took place after the first second, when the two models coincide.

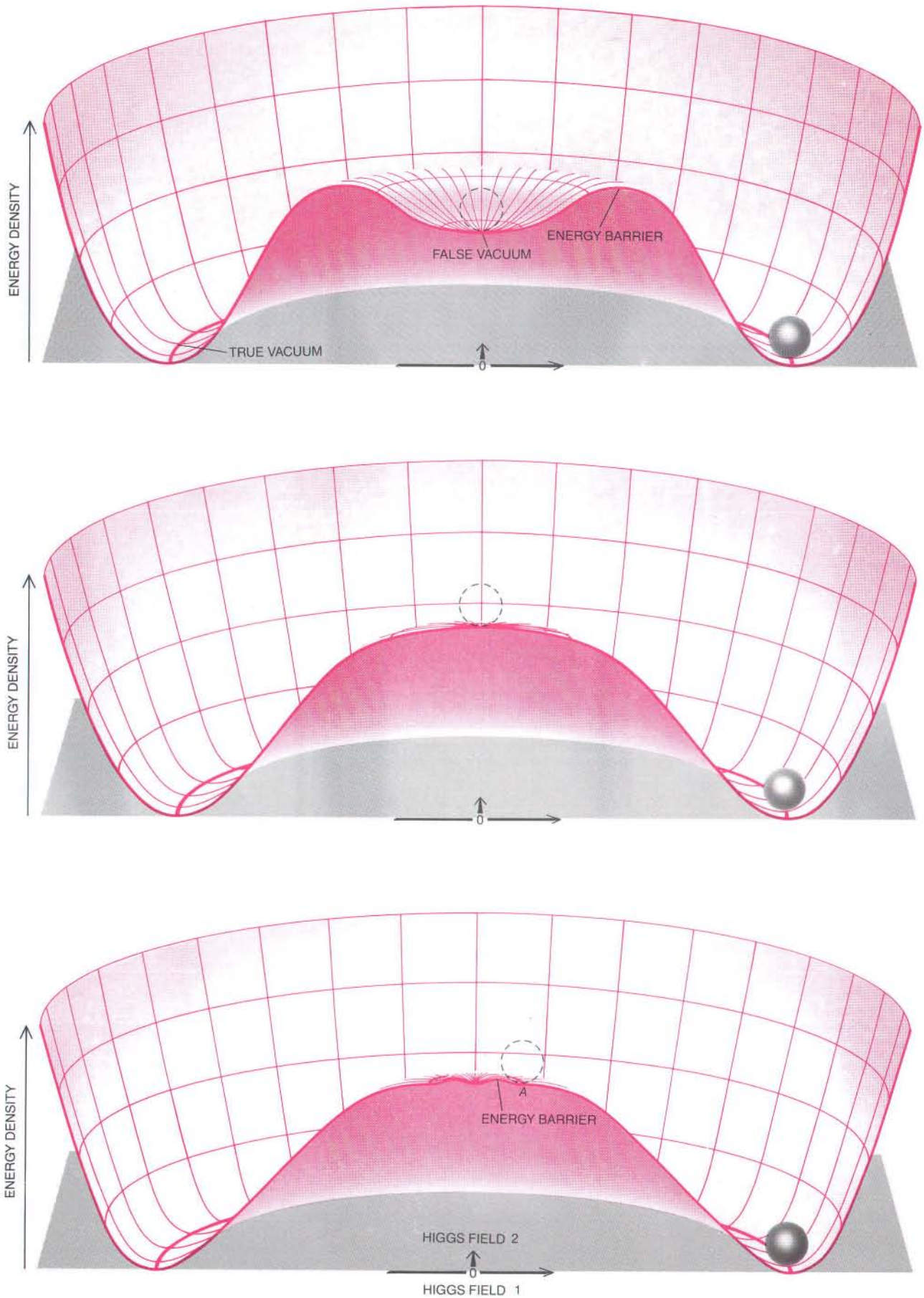
Until about five years ago there were few serious attempts to describe the universe during its first second. The temperature in this period is thought to have been higher than 10 billion kelvins, and little was known about the properties of matter under such conditions. Relying on recent developments in the physics of elementary particles, however, cosmologists are now attempting to understand the history of the universe back to 10^{-45} second after its beginning. (At even earlier times the energy density would have been so great that Einstein's general theory of

ALAN H. GUTH and PAUL J. STEINHARDT are physicists who share an interest in the early history of the universe and particularly the first 10^{-45} second of that history. Guth attended the Massachusetts Institute of Technology as an undergraduate and a graduate student; his Ph.D. in physics was awarded by M.I.T. in 1972. He writes: "I held postdoctoral positions at Princeton University, Columbia University, Cornell University and the Stanford Linear Accelerator Center (SLAC). During most of that period of years, I worked on rather abstract mathematical problems in elementary particle theory and knew no more about developments in cosmology than the average layman does. While I was at Cornell, however, Henry Tye, a fellow postdoctoral worker, persuaded me (with great difficulty) to join him in studying the production of magnetic monopoles in the early universe, and that was how my career changed direction. I continued the work in the following academic year at SLAC. Shortly thereafter I returned to M.I.T." Guth is now professor of physics there. Steinhardt was graduated from the California Institute of Technology with a B.S. in 1974. His M.A. (1975) and Ph.D. (1978) in physics are from Harvard University. From 1979 to 1981 he was a junior fellow in the Society of Fellows at Harvard. In 1981 he moved to the University of Pennsylvania, where he is professor of physics.

INFLATIONARY MODEL of the universe is represented by the colored curves in this set of graphs, showing how several properties of the observed universe could have changed with time starting at 10^{-45} second after the big bang. The gray curves represent the standard big bang model, which is coincident with the inflationary model for all times after 10^{-30} second. For comparison the graph of temperature (*top*) also includes an indication of the boiling point of water (373 kelvins) and the temperature at the center of a typical star (10 million kelvins). Similarly the graph of energy density (*middle*) indicates the energy density of water (10^{21} ergs per cubic centimeter) and of an atomic nucleus (10^{36} ergs per cubic centimeter). On the graph of spatial dimensions (*bottom*) each cosmological model is represented by two curves. One curve shows the region of space that evolves to become the observed universe, and the other shows the horizon distance: the total distance a light signal could have traveled since the beginning of the universe. One problem of the standard model, known as the horizon problem, arises from the fact that the horizon distance is much smaller than the radius of the observable universe for most of its history. In the inflationary model the horizon distance greatly exceeds the radius of the observable universe at all times. On the time axis, several significant events are marked. *A* indicates the time of the phase transition predicted in the standard big bang model by grand unified theories of the interactions of elementary particles; at the high temperature prevailing before this time, the various nongravitational forces acting between particles are thought to have been related to one another by a symmetry that was spontaneously broken when the temperature fell to a critical value of about 10^{27} kelvins. A key feature of the inflationary model is the prolongation of the phase transition, which extends through a period called the inflationary era (*color band*); during this era the universe expands by an extraordinary factor, perhaps 10^{50} or more. Meanwhile the temperature plunges, but it is stabilized at about 10^{22} kelvins by quantum effects that arise in the context of general relativity. The gray band labeled *B* indicates the period when the lightest atomic nuclei were formed, and *C* indicates the time when the universe became transparent to electromagnetic radiation.

Reprinted from the May 1984 issue


TIME (YEARS)
10^{-50}
10^{-40}
10^{-30}
10^{-20}
10^{-10}
1
10^{10}
TEMPERATURE (DEGREES KELVIN)
10^{30}
10^{20}
10^{10}
1
STANDARD MODEL
INFLATIONARY MODEL
STELLAR INTERIOR
BOILING WATER
ENERGY DENSITY (ERGS PER CUBIC CENTIMETER)
10^{120}
10^{90}
10^{60}
10^{30}
1
STANDARD MODEL
INFLATIONARY MODEL
ATOMIC NUCLEUS
WATER
DISTANCE (CENTIMETERS)
10^{60}
10^{50}
10^{40}
10^{30}
10^{20}
10^{10}
10^{-10}
10^{-20}
10^{-30}
10^{-40}
10^{-50}
10^{-60}
HORIZON DISTANCE (INFLATIONARY MODEL)
RADIUS OF OBSERVABLE UNIVERSE (STANDARD MODEL)
HORIZON DISTANCE (STANDARD MODEL)
A
RADIUS OF OBSERVABLE UNIVERSE (INFLATIONARY MODEL)
INFLATIONARY ERA
B
C
10^{-45}
10^{-35}
10^{-25}
10^{-15}
10^{-5}
10^{5}
10^{15}
TIME (SECONDS)
PRESENT


ENERGY DENSITY
FALSE VACUUM
ENERGY BARRIER
TRUE VACUUM
0
ENERGY DENSITY
0
ENERGY DENSITY
A
ENERGY BARRIER
HIGGS FIELD 2
0
HIGGS FIELD 1

relativity would have to be replaced by a quantum theory of gravity, which so far does not exist.) When the standard big bang model is extended to these earlier times, various problems arise. First, it becomes clear that the model requires a number of stringent, unexplained assumptions about the initial conditions of the universe. In addition most of the new theories of elementary particles imply that the standard model would lead to a tremendous overproduction of the exotic particles called magnetic monopoles. (Each such monopole corresponds to an isolated north or south magnetic pole.)

The inflationary universe was invented to overcome these problems. The equations that describe the period of inflation have a very attractive feature: from almost any initial conditions the universe evolves to precisely the state that had to be assumed as the initial one in the standard model. Moreover, the predicted density of magnetic monopoles becomes small enough to be consistent with observations. In the context of the recent developments in elementary particle theory, the inflationary model seems to be a natural solution to many of the problems of the standard big bang picture.

The standard big bang model is based on several assumptions. First, it is assumed that the fundamental laws of physics do not change with time and that the effects of gravitation are correctly described by Einstein's general theory of relativity. It is also assumed that the early universe was filled with an almost perfectly uniform, expanding, intensely hot gas of elementary particles in thermal equilibrium. The gas filled all of space, and the gas and space expanded together at the same rate. When they are averaged over large regions, the densities of matter and energy have remained nearly uniform from place to place as the universe has evolved. It is further assumed that any changes in the state of the matter and the radiation have been so smooth that they have had a negligible effect on the thermodynamic history of the universe. The violation of the last assumption is a key to the inflationary universe model.

The big bang model leads to three important, experimentally testable predictions. First, the model predicts that as the universe expands, galaxies recede from one another with a velocity proportional to the distance between them. In the 1920s Edwin P. Hubble inferred just such an expansion law from his study of the redshifts of distant galaxies. Second, the big bang model predicts that there should be a background of microwave radiation bathing the universe as a remnant of the intense heat of its origin. The universe became transparent to this radiation several hundred thousand years after the big bang. Ever since then the matter has been clumping into stars, galaxies and the like, but the radiation has simply continued to expand and redshift—in effect to cool. In 1964 Arno A. Penzias and Robert W. Wilson of Bell Telephone Laboratories discovered a background of microwave radiation received uniformly from all directions with an effective temperature of about three kelvins. Third, the model leads to successful predictions of the formation of light atomic nuclei from protons and neutrons during the first minutes after the big bang. Successful predictions can be obtained in this way for the abundance of helium 4, deuterium, helium 3 and lithium 7. (Heavier nuclei are thought to have been produced much later in the interior of stars.)

Unlike the successes of the big bang model, all of which pertain to events a second or more after the big bang, the problems all concern times when the universe was much less than a second old. One set of problems has to do with the special conditions the model requires as the universe emerged from the big bang.

The first problem is the difficulty of explaining the large-scale uniformity of the observed universe. The large-scale uniformity is most evident in the microwave background radiation, which is known to be uniform in temperature to about one part in 10,000. In the standard model the universe evolves much too quickly to allow this unifor-

ENERGY DENSITY of the universe is represented in these three-dimensional diagrams as a function of two Higgs fields, members of a special set of fields postulated in grand unified theories to account for spontaneous symmetry breaking. Each surface shown in cross section is rotationally symmetric about a vertical axis, which corresponds to a state in which both Higgs fields have a value of zero. In the absence of thermal excitations, this state of unbroken symmetry, known as the false vacuum, would have an energy density of about 10^{95} ergs per cubic centimeter, or some 10^{59} times the energy density of an atomic nucleus. The rotational symmetry is broken whenever one of the Higgs fields acquires a nonzero value (or both of them do). Here the theory has been formulated in such a way that the states of lowest energy density, known as the true-vacuum states, are states of broken symmetry, forming a circle in the horizontal plane at the bottom of each diagram. In this analogy the evolution of the universe can be traced by imagining a ball rolling on the surface. The ball's distance from the central axis represents the combined values of the Higgs fields, and its height above the horizontal surface represents the energy density of the universe. When the Higgs fields both have a value of zero, the ball is poised at the axis of symmetry; when the Higgs fields have a value that corresponds to the lowest possible energy density, the ball is lying somewhere in the trough that defines the broken-symmetry, or true-vacuum, states. In the original form of the inflationary universe model, it was assumed that the energy density function had the shape in the diagram at the top. The inflationary episode would then take place while the universe was in the false-vacuum state. If the laws of classical physics applied, this state would be absolutely stable, because there would be no energy available to carry the Higgs fields over the intervening energy barrier. According to the laws of quantum physics, however, the fields in small regions of space can "tunnel" through the energy barrier, forming bubbles of the broken-symmetry phase, which would then start to grow. In the new inflationary model (*middle diagram*) there is no energy barrier; instead the false vacuum is at the top of a rather flat plateau. Under these circumstances, the transition from the false vacuum to the broken-symmetry phase occurs by means of a slow-rollover mechanism: the Higgs fields are pushed from their initial value of zero by thermal or quantum fluctuations, and they proceed toward their true-vacuum values much as a ball would roll over a plateau of the same shape. The accelerated expansion of the universe takes place during the early stages of the rollover, while the energy density remains roughly constant. A single domain of broken-symmetry phase could then grow large enough to encompass the entire observable universe. When the Higgs fields reached the bottom of the trough, they would oscillate about the lowest energy density value, causing a reheating of the universe. In a variant of the new inflationary model (*bottom diagram*), the false vacuum is surrounded by a small energy barrier. As in the original inflationary model, the false vacuum decays by the random formation of bubbles, created by the tunneling of the Higgs fields through the energy barrier. Because the energy barrier is small in this case, the Higgs fields tunnel only as far as the circle labeled *A*. Since the slope is quite flat at *A*, the Higgs fields evolve very slowly toward their true-vacuum values. The accelerated expansion of the universe continues as long as the Higgs fields remain near *A*, and a single bubble could grow large enough to encompass the observable universe.

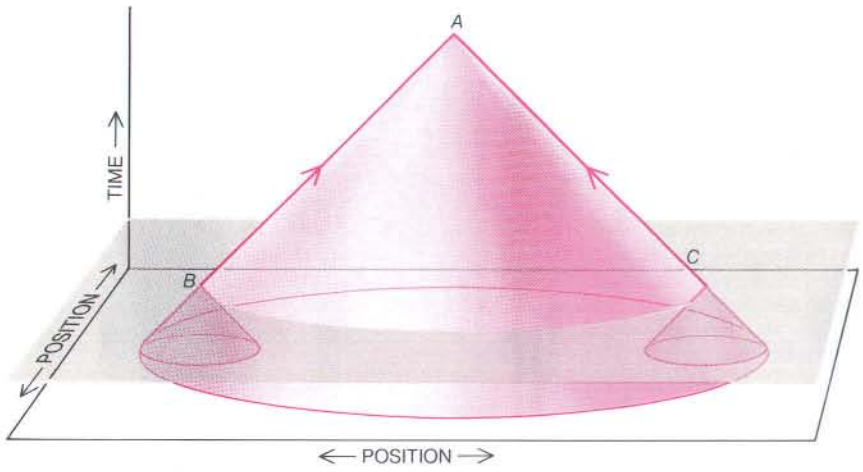
mity to be achieved by the usual processes whereby a system approaches thermal equilibrium. The reason is that no information or physical process can propagate faster than a light signal. At any given time there is a maximum distance, known as the horizon distance, that a light signal could have traveled since the beginning of the universe. In the standard model the sources of the microwave background radiation observed from opposite directions in the sky were separated from each other by more than 90 times the horizon distance when the radiation was emitted. Since the regions could not have communicated, it is difficult to see how they could have evolved conditions so nearly identical.

The puzzle of explaining why the universe appears to be uniform over distances that are large compared with the horizon distance is known as the horizon problem: It is not a genuine inconsistency of the standard model; if the uniformity is assumed in the initial conditions, the universe will evolve uniformly. The problem is that one of the most salient features of the observed universe—its large-scale uniformity—cannot be explained by the standard model; it must be assumed as an initial condition.

Even with the assumption of large-scale uniformity, the standard big bang model requires yet another assumption to explain the nonuniformity observed on smaller scales. To account for the clumping of matter into galaxies, clusters of galaxies, superclusters of clusters and so on, a spectrum of primordial inhomogeneities must be assumed as part of the initial conditions. The fact that the spectrum of inhomogeneities has no explanation is a drawback in itself, but the problem becomes even more pronounced when the model is extended back to 10^{-45} second after the big bang. The incipient clumps of matter develop rapidly with time as a result of their gravitational self-attraction, and so a model that begins at a very early time must begin with very small inhomogeneities. To begin at 10^{-45} second, the matter must start in a peculiar state of extraordinary but not quite perfect uniformity. A normal gas in thermal equilibrium would be far too inhomogeneous, because of the random motion of particles. This peculiarity of the initial state of matter required by the standard model is called the smoothness problem.

Another subtle problem of the standard model concerns the energy density of the universe. According to general relativity, the space of the universe can in principle be curved, and the nature of the curvature depends on the energy density. If the energy density exceeds a certain critical value, which depends on the expansion rate, the universe is said to be closed: space curves back on itself to form a finite volume with no boundary. (A familiar analogy is the surface of a sphere, which is finite in area and has no boundary.) If the energy density is less than the critical density, the universe is open: space curves but does not turn back on itself, and the volume is infinite. If the energy density is just equal to the critical density, the universe is flat: space is described by the familiar Euclidean geometry (again with infinite volume).

The ratio of the energy density of the universe to the critical density is a quantity cosmologists designate by the Greek letter Ω (omega). The value $\Omega = 1$ (corresponding to a flat universe) represents a state of unstable equilibrium. If Ω was ever exactly equal to 1, it would remain exactly equal to 1 forever. If Ω differed slightly from 1 an instant after the big bang, however, the deviation from 1 would grow rapidly with time. Given this instability, it is surprising that Ω is measured today as being between 0.1 and 2. (Cosmologists are still not sure whether the universe is open, closed or flat.) In order for Ω to be in this rather narrow range today, its value a second after the big bang had to equal 1 to within one part in 10^{15}. The standard model offers no explanation of why Ω began so close to 1 but merely assumes the fact as an initial condition. This shortcoming of the standard model, called the flatness problem, was first pointed out in 1979 by Robert H. Dicke and P. James E. Peebles of Princeton University.



HORIZON PROBLEM is a serious drawback of the standard big bang theory. In this three-dimensional space-time diagram the scales have been drawn in a nonlinear way so that the trajectory of a light pulse is represented by a line at 45 degrees to the vertical axis. Our position in space and time is indicated by the point *A*. Since no signal can travel faster than the speed of light, we can receive signals only from the colored region, called our past light cone. Events outside the past light cone of a given point cannot influence an event at that point in any way. The gray horizontal plane shows the time at which the microwave background radiation was released. Radiation that is now reaching us from opposite directions was released at points *B* and *C*, and since then it has traveled along our past light cone to point *A*. The past light cone of point *B* has no intersection with the past light cone of point *C*, and therefore the two points were not subject to any common influences. The horizon problem is the difficulty of explaining how the radiation received from the two opposite directions came to be at the same temperature. In the standard model the large-scale uniformity of temperature evident in the microwave background radiation must be assumed as an initial condition of the universe.

The successes and drawbacks of the big bang model we have considered so far involve cosmology, astrophysics and nuclear physics. As the big bang model is traced backward in time, however, one reaches an epoch for which these branches of physics are no longer adequate. In this epoch all matter is decomposed into its elementary particle constituents. In an attempt to understand this epoch, cosmologists have made use of recent

progress in the theory of elementary particles. Indeed, one of the important developments of the past decade has been the fusing of interests in particle physics, astrophysics and cosmology. The result for the big bang model appears to be at least one more success and at least one more failure.

Perhaps the most important development in the theory of elementary particles over the past decade has been the notion of grand unified theories, the prototype of which was proposed in 1974 by Howard M. Georgi and Sheldon Lee Glashow of Harvard University. The theories are difficult to verify experimentally because their most distinctive predictions apply to energies far higher than the energies that can be reached with today's particle accelerators. Nevertheless, the theories have some experimental support, and they unify the understanding of elementary particle interactions so elegantly that many physicists find them extremely attractive.

The basic idea of a grand unified theory is that what were perceived to be three independent forces—the strong, the weak and the electromagnetic—are actually parts of a single unified force. In the theory a symmetry relates one force to another. Since experimentally the forces are very different in strength and character, the theory is constructed so that the symmetry is spontaneously broken in the present universe.

A spontaneously broken symmetry is one that is present in the underlying theory describing a system but is hidden in the equilibrium state of the system. For example, a liquid described by physical laws that are rotationally symmetric is itself rotationally symmetric: the distribution of molecules looks the same no matter how the liquid is turned. When the liquid freezes into a crystal, however, the atoms arrange themselves along crystallographic axes, and the rotational symmetry is broken. One would expect that if the temperature of a system in a broken-symmetry state were raised, it could undergo a kind of phase transition to a state in which the symmetry is restored, just as a crystal can melt into a liquid. Grand unified theories predict such a transition at a critical temperature of roughly 10^{27} kelvins.

One novel property of the grand unified theories has to do with the particles called baryons, a class whose most important members are the proton and the neutron. In all physical processes observed up to now, the number of baryons minus the number of antibaryons does not change; in the language of particle physics, the total baryon number of the system is said to be conserved. A consequence of such a conservation law is that the proton must be absolutely stable; because it is the lightest baryon, it cannot decay into another particle without changing the total baryon number. Experimentally the lifetime of the proton is known to exceed 10^{31} years.

Grand unified theories imply that baryon number is not exactly conserved. At low temperature, in the broken-symmetry phase, the conservation law is an excellent approximation, and the observed limit on the proton lifetime is consistent with at least many versions of grand unified theories. At high temperature, however, processes that change the baryon number of a system of particles are expected to be quite common.

TYPE OF UNIVERSE	RATIO OF ENERGY DENSITY TO CRITICAL DENSITY (Ω)	SPATIAL GEOMETRY	VOLUME	TEMPORAL EVOLUTION
CLOSED	>1	POSITIVE CURVATURE (SPHERICAL)	FINITE	EXPANDS AND RECOLLAPSES
OPEN	<1	NEGATIVE CURVATURE (HYPERBOLIC)	INFINITE	EXPANDS FOREVER
FLAT	1	ZERO CURVATURE (EUCLIDEAN)	INFINITE	EXPANDS FOREVER, BUT EXPANSION RATE APPROACHES ZERO

THREE TYPES OF UNIVERSE, classified as closed, open and flat, can arise from the standard big bang model (under the usual assumption that the equations of general relativity are not modified by the addition of a cosmological term). The distinction between the different geometries depends on the quantity designated Ω, the ratio of the energy density of the universe to some critical density, whose value depends in turn on the rate of expansion of the universe. The value of Ω today is known to lie between 0.1 and 2, which implies that its value a second after the big bang was equal to 1 to within one part in 10^{15}. The failure of the standard big bang model to explain why Ω began so close to 1 is called the flatness problem.

One direct result of combining the big bang model with grand unified theories is the successful prediction of the asymmetry of matter and antimatter in the universe. It is thought that all the stars, galaxies and dust observed in the universe are in the form of matter rather than antimatter; their nuclear particles are baryons rather than antibaryons. It follows that the total baryon number of the observed universe is about 10^{78}. Before the advent of grand unified theories, when baryon number was thought to be conserved, this net baryon number had to be postulated as yet another initial condition of the universe. When grand unified theories and the big bang picture are combined, however, the observed excess of matter over antimatter can be produced naturally by elementary particle interactions at temperatures just below the critical temperature of the phase transition. Calculations in the grand unified theories depend on too many arbitrary parameters for a quantitative prediction to be made, but the observed matter-antimatter asymmetry can be produced with a reasonable choice of values for the parameters.

A serious problem that results from combining grand unified theories with the big bang picture is that a large number of defects are generally formed during the transition from the symmetric phase to the broken-symmetry phase. The defects are created when regions of symmetric phase undergo a transition to different broken-symmetry states. In an analogous situation, when a liquid crystallizes, different regions may begin to crystallize with different orientations of the crystallographic axes. The domains of different crystal orientation grow and coalesce, and it is energetically favorable for them to smooth the misalignment along their boundaries. The smoothing is often imperfect, however, and localized defects remain.

In the grand unified theories there are serious cosmological problems associated with pointlike defects, which correspond to magnetic monopoles, and surfacelike defects, called domain walls. Both are expected to be extremely stable and extremely massive. (The monopole can be shown to be about 10^{16} times as heavy as the proton.) A

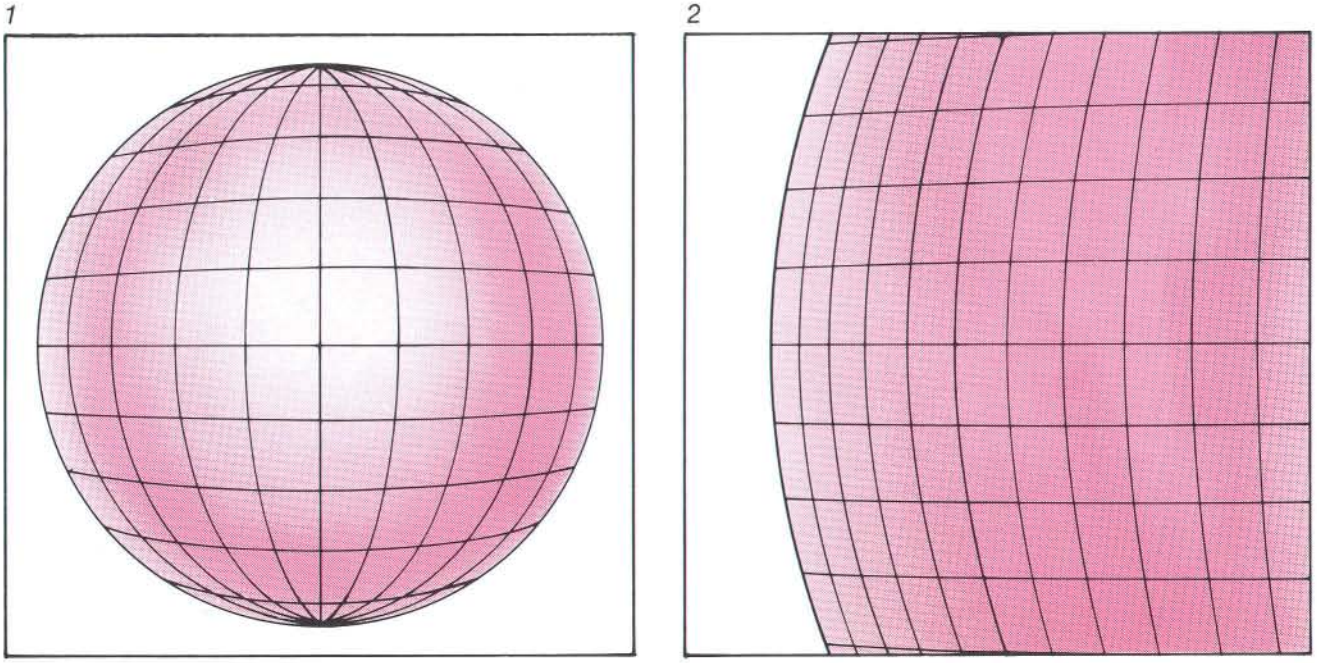
domain of correlated broken-symmetry phase cannot be much larger than the horizon distance at that time, and so the minimum number of defects created during the transition can be estimated. The result is that there would be so many defects after the transition that their mass would dominate the energy density of the universe and thereby speed up its subsequent evolution. The cosmic microwave background radiation would reach its present temperature of three kelvins only 30,000 years after the big bang instead of 10 billion years, and all the successful predictions of the big bang model would be lost. Thus, any successful union of grand unified theories and the big bang picture must incorporate some mechanism to drastically suppress the production of magnetic monopoles and domain walls.

The inflationary universe model appears to provide a satisfactory solution to these problems. Before the model can be described, however, we must first explain a few more of the details of symmetry breaking and phase transitions in grand unified theories.

All modern particle theories, including the grand unified theories, are examples of quantum field theories. The best-known field theory is the one that describes electromagnetism. According to the classical (nonquantum) theory of electromagnetism developed by James Clerk Maxwell in the 1860s, electric and magnetic fields have a well-defined value at every point in space, and their variation with time is described by a definite set of equations. Maxwell's theory was modified early in the 20th century in order to achieve consistency with the quantum theory. In the classical theory, it is possible to increase the energy of an electromagnetic field by any amount; in the quantum theory, however, the increases in energy can come only in discrete lumps, the quanta, which in this case are called photons. The photons have both wavelike and particlelike properties, but in the lexicon of modern physics they are usually called particles. In general the formulation of a quantum field theory begins with a classical theory of fields, and it becomes a theory of particles when the rules of the quantum theory are applied.

As we have already mentioned, an essential ingredient of grand unified theories is the phenomenon of spontaneous symmetry breaking. The detailed mechanism of spontaneous symmetry breaking in grand unified theories is simpler in many ways than the analogous mechanism in crystals. In a grand unified theory, spontaneous symmetry breaking is accomplished by including in the formulation of the theory a special set of fields known as Higgs fields (after Peter W. Higgs of the University of Edinburgh). The symmetry is unbroken when all the Higgs fields have a value of zero, but it is spontaneously broken whenever at least one of the Higgs fields acquires a nonzero value. Furthermore, it is possible to formulate the theory in such a way that a Higgs field has a nonzero value in the state of lowest energy density, which in this context is known as the true vacuum. At temperatures greater than about 10^{27} kelvins, thermal fluctuations drive the equilibrium value of the Higgs field to zero, resulting in a transition to the symmetric phase.

We have now assembled enough background information to describe the inflationary model of the universe, beginning with the form in which it was first proposed by one of us (Guth) in 1980. Any cosmological model must begin with some assumptions about the initial conditions, but for the inflationary model the initial conditions can be rather arbitrary. One must assume, however, that the early universe included at least some regions of gas that were hot compared with the critical temperature of the phase transition and that were also expanding. In such a hot region the Higgs field would have a value of zero. As the expansion caused the temperature to fall, it would become thermodynamically favorable for



SOLUTION OF THE FLATNESS PROBLEM is illustrated by these drawings of an inflating sphere. The illustration shows how a flat spatial geometry (which corresponds to a value of Ω equal to 1) can be produced by the inflationary scenario

the Higgs field to acquire a nonzero value, bringing the system to its broken-symmetry phase.

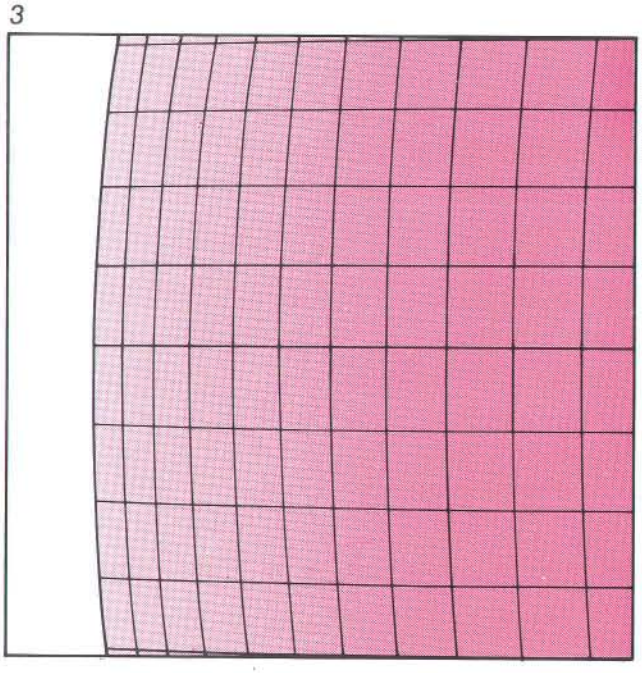
For some values of the unknown parameters of the grand unified theories, this phase transition would occur very slowly compared with the cooling rate. As a result, the system could cool to well below 10^{27} kelvins with the value of the Higgs field remaining at zero. This phenomenon, known as supercooling, is quite common in condensed-matter physics; water, for example, can be supercooled to more than 20 degrees below its freezing point, and glasses are formed by rapidly supercooling a liquid to a temperature well below its freezing point.

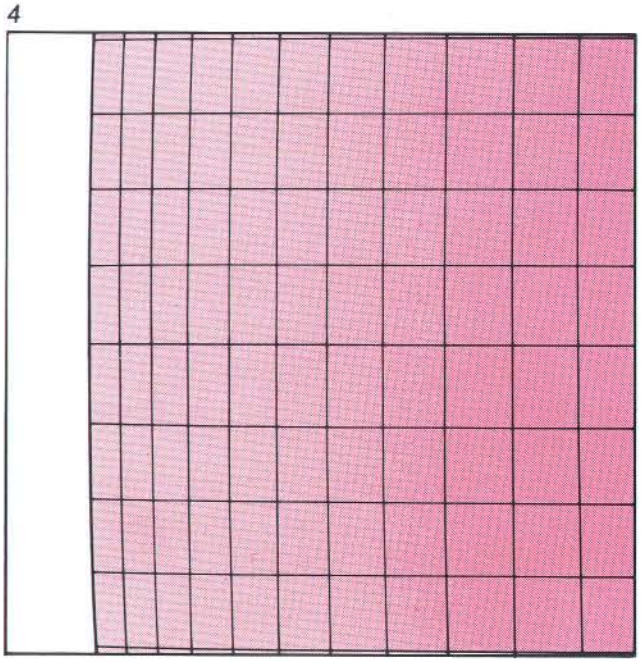
As the region of gas continued to supercool, it would approach a peculiar state of matter known as a false vacuum. This state of matter has never been observed, but it has properties that are unambiguously predicted by quantum field theory. The temperature, and hence the thermal component of the energy density, would rapidly decrease, and the energy density of the state would be concentrated entirely in the Higgs field. A zero value for the Higgs field implies a large energy density for the false vacuum. In the classical form of the theory, such a state would be absolutely stable, even though it would not be the state of lowest energy density. States with a lower energy density would be separated from the false vacuum by an intervening energy barrier, and there would be no energy available to take the Higgs field over the barrier [*see top illustration on page 50*].

In the quantum version of the model the false vacuum is not absolutely stable. Under the rules of the quantum theory, all the fields would be continually fluctuating. As was first described by Sidney R. Coleman of Harvard, a quantum fluctuation would occasionally cause the Higgs field in a small region of space to "tunnel" through the energy barrier, nucleating a "bubble" of the broken-symmetry phase. The bubble would then start to grow at a speed that would rapidly approach the speed of light, converting the false vacuum into the broken-symmetry phase. The rate at which bubbles form depends sensitively on the unknown parameters of the grand unified theory; in the inflationary model it is assumed that the rate would be extremely low.

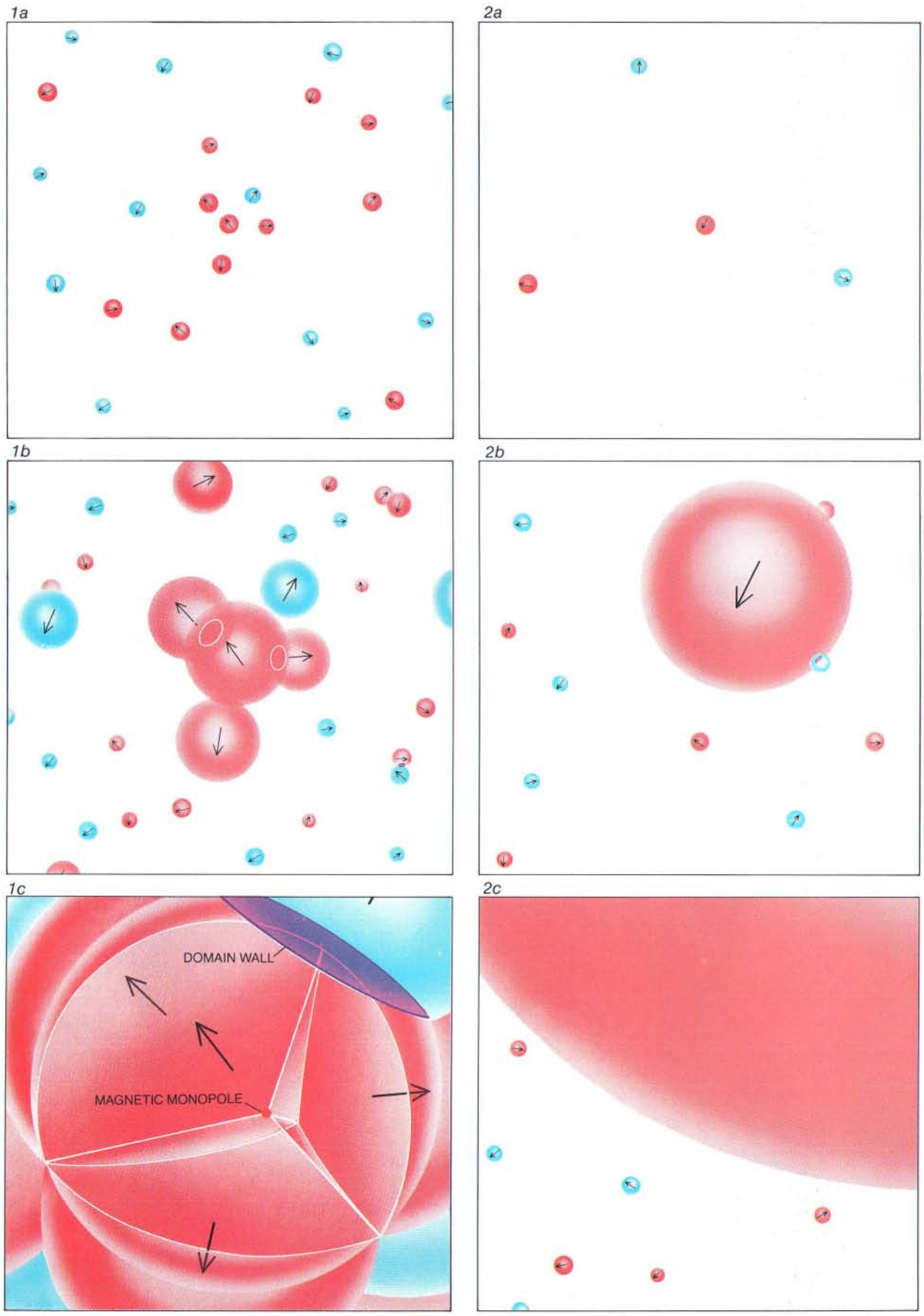
The most peculiar property of the false vacuum is probably its pressure, which is both large and negative. To understand why this is so, consider again the process by which a bubble of true vacuum would grow into a region of false vacuum. The growth is favored energetically because the true vacuum has a lower energy density than the false vacuum. The growth also indicates, however, that the pressure of the true vacuum must be higher than the pressure of the false vacuum, forcing the bubble wall to grow outward. Because the pressure of the true vacuum is zero, the pressure of the false vacuum must be negative. A more detailed argument shows that the pressure of the false vacuum is equal to the negative value of its energy density (when the two quantities are measured in the same units).

The negative pressure would not result in mechanical forces within the false vacuum, because mechanical forces arise only from differences in pressure. Nevertheless, there would be gravitational effects. Under ordinary circumstances, the expansion of the region of gas would be slowed by the mutual gravitational attraction of the matter within it. In Newtonian physics this attraction is proportional to the mass density, which in relativistic theories is equal to the energy density divided by the square of the speed of light. According to general relativity, the pressure also contributes to the attraction; to be specific, the gravitational force is proportional to the energy density plus three times the pressure. For the false vacuum, the contribution made by the pressure would overwhelm the energy density contri-





in a simple and natural way. In each successive frame the sphere is inflated by a factor of three (and the number of grid lines on the surface is increased by three). The surface curvature quickly becomes undetectable at this scale.


1a
2a
1b
2b
1c
2c
DOMAIN WALL
MAGNETIC MONOPOLE

bution and would have the opposite sign. Therefore, the bizarre notion of negative pressure leads to the even more bizarre effect of a gravitational force that is effectively repulsive. As a result, the expansion of the region would be accelerated and the region would grow exponentially, doubling in diameter during each interval of approximately 10^{-34} second.

This period of accelerated expansion is called the inflationary era, and it is the key element of the inflationary model of the universe. According to the model, the inflationary era continued for 10^{-32} second or longer, and during this period the diameter of the universe increased by a factor of 10^{50} or more. It is assumed that after this colossal expansion the transition to the broken-symmetry phase finally took place. The energy density of the false vacuum was then released, resulting in a tremendous amount of particle production. The region was reheated to a temperature of almost 10^{27} kelvins. (In the language of thermodynamics the energy released is called the latent heat; it is analogous to the energy released when water freezes.) From this point onward, the region would continue to expand and cool at the rate described by the standard big bang model. A volume the size of the observable universe would lie well within such a region.

The horizon problem is avoided in a straightforward way. In the inflationary model the observed universe evolves from a region that is much smaller in diameter (by a factor of 10^{50} or more) than the corresponding region in the standard model. Before inflation begins, the region is much smaller than the horizon distance, and it has time to homogenize and reach thermal equilibrium. This small homogeneous region is then inflated to become large enough to encompass the observed universe. Thus, the sources of the microwave background radiation arriving today from all directions in the sky were once in close contact; they had time to reach a common temperature before the inflationary era began.

The flatness problem is also evaded in a simple and natural way. The equations describing the evolution of the universe during the inflationary era are different from those for the standard model, and it turns out that the ratio Ω is driven rapidly toward 1, no matter what value it had before inflation. This behavior is most easily understood by recalling that a value of $\Omega = 1$ corresponds to a space that is geometrically flat. The rapid expansion causes the space to become flatter just as the surface of a balloon becomes flatter when it is inflated. The mechanism driving Ω toward 1 is so effective that one is led to an almost rigorous prediction: the value of Ω today should be very accurately equal to 1. Many astronomers (although not all) think a value of 1 is consistent with current observations, but a more reliable determination of Ω would provide a crucial test of the inflationary model.

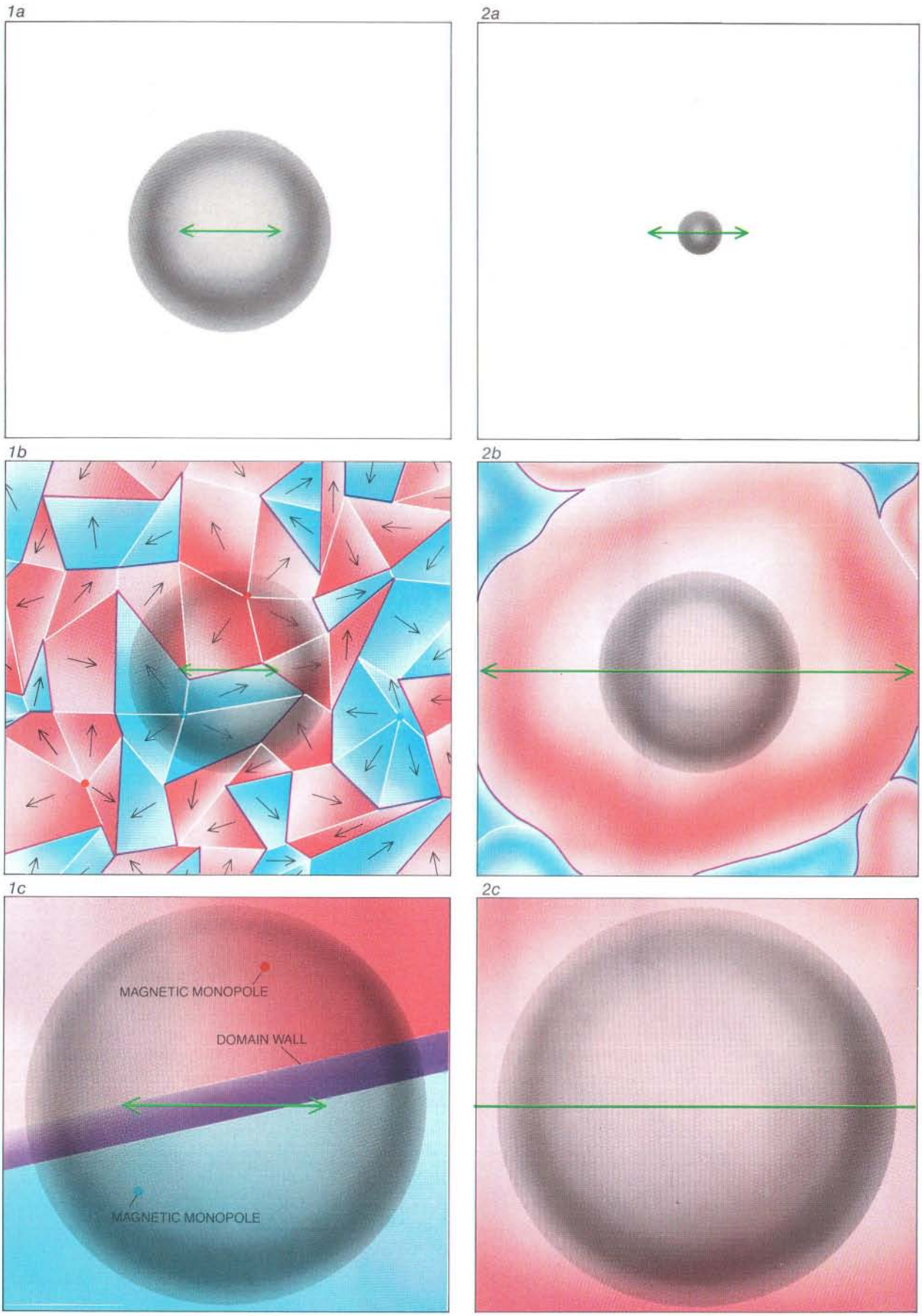
In the form in which the inflationary model was originally proposed, it had a crucial flaw: under the circumstances described, the phase transition itself would create inhomogeneities much more extreme than those observed today. As we have already described, the phase transition would take place by the random nucleation of bubbles of the new phase. It can be shown that the bubbles would always remain in finite clusters disconnected from one another and that each cluster would be dominated by a single largest bubble. Almost all the energy in the cluster would be initially concentrated in the surface of the largest bubble, and there is no apparent mechanism to redistribute energy uniformly. Such a configuration bears no resemblance to the observed universe.

EXPANDING BUBBLES of broken-symmetry phase form in an expanding region of symmetric phase in the two highly schematic time sequences on the opposite page. The sequence representing the standard big bang model (*left*) covers a much shorter time span than the sequence representing the original form of the inflationary model (*right*). In both cases the Higgs fields have a value of zero in the region outside the bubbles, whereas at least one Higgs field has a nonzero value inside each bubble. In a grand unified theory the broken-symmetry states can in general be distinguished by parameters of two kinds: discrete and continuous. Here each bubble is labeled in two ways: by a color (*blue or red*) to indicate the discrete parameter and by an internal black arrow to indicate the value of the continuous parameter. In the standard model the bubbles would quickly coalesce and complete the transition from the symmetric phase to the broken-symmetry phase. A surface-like defect called a domain wall would form at any boundary between regions with different values of the discrete parameter (*purple areas*). Within a region of uniform color, a pointlike defect called a magnetic monopole would form at a center created by the intersection of many bubbles whenever the arrow representing the continuous parameter points everywhere away from the center. In the original form of the inflationary model, the rapid expansion of the false-vacuum, or symmetric-phase, region would keep the bubbles from ever coalescing. Either hypothetical situation has consequences that are contrary to observation; the new inflationary model was developed in order to avoid both of them.

For almost two years after the invention of the inflationary universe model, it remained a tantalizing but clearly imperfect solution to a number of important cosmological problems. Near the end of 1981, however, a new approach was developed by A. D. Linde of the P. N. Lebedev Physical Institute in Moscow and independently by Andreas Albrecht and one of us (Steinhardt) of the University of Pennsylvania. This approach, known as the new inflationary universe, avoids all the problems of the original model while maintaining all its successes.

The key to the new approach is to consider a special form of the energy density function that describes the Higgs field [*see middle illustration on page 50*]. Quantum field theories with energy density functions of this type were first studied by Coleman, working in collaboration with Erick J. Weinberg of Columbia University. In contrast to the more typical case shown in the top illustration on page 50, there is no energy barrier separating the false vacuum from the true vacuum; instead the false vacuum lies at the top of a rather flat plateau. In the context of grand unified theories, such an energy density function is achieved by a special choice of parameters. As we shall explain below, this energy density function leads to a special type of phase transition that is sometimes called a slow-rollover transition.

The scenario begins just as it does in the original inflationary model. Again one must assume the early universe had regions that were hotter than about 10^{27} kelvins and were also expanding. In these regions, thermal fluctuations would drive the equilibrium value of the Higgs fields to zero, and the symmetry would be unbroken. As the temperature fell, it would become thermodynamically favorable for the system to undergo a phase transition in which at least one of the Higgs fields acquired a nonzero value, resulting in a broken-symmetry phase. As in the previous case, however, the rate of this phase transition would be extremely low compared with the rate of cooling. The system would supercool to a negli-


1a
2a
1b
2b
1c
2c
MAGNETIC MONOPOLE
DOMAIN WALL
MAGNETIC MONOPOLE

gible temperature with the Higgs field remaining at zero, and the resulting state would again be considered a false vacuum.

The important difference in the new approach is the way in which the phase transition would take place. Quantum fluctuations or small residual thermal fluctuations would cause the Higgs field to deviate from zero. In the absence of an energy barrier the value of the Higgs field would begin to increase steadily; the rate of increase would be much like that of a ball rolling down a hill of the same shape as the curve of the energy density function, under the influence of a frictional drag force. Since the energy density curve is almost flat near the point where the Higgs field vanishes, the early stage of the evolution would be very slow. As long as the Higgs field remained close to zero, the energy density would be almost the same as it is in the false vacuum. As in the original scenario, the region would undergo accelerated expansion, doubling in diameter every 10^{-34} second or so. In this case, however, the expansion would cease to accelerate when the value of the Higgs field reached the steeper part of the curve. By computing the time required for the Higgs field to evolve, the amount of inflation can be determined. An expansion factor of 10^{50} or more is quite plausible, but the actual factor depends on the details of the particle theory one adopts.

So far the description of the phase transition has been slightly oversimplified. There are actually many different states of broken symmetry, just as there are many possible orientations for the axes of a crystal. There are a number of Higgs fields, and the various broken-symmetry states are distinguished by the combination of Higgs fields that acquire nonzero values. Since the fluctuations that drive the Higgs fields from zero are random, different regions of the primordial universe would be driven toward different broken-symmetry states, each region forming a domain with an initial radius of roughly the horizon distance. At the start of the phase transition the horizon distance would be about 10^{-24} centimeter. Once the domain formed, with the Higgs fields deviating slightly from zero in a definite combination, it would evolve toward one of the stable broken-symmetry states and would inflate by a factor of 10^{50} or more. The size of the domain after inflation would then be greater than 10^{26} centimeters. The entire observable universe, which at that time would be only about 10 centimeters across, would be able to fit deep inside a single domain.

In the course of this enormous inflation, any density of particles that might have been present initially would be diluted to virtually zero. The energy content of the region would then consist entirely of the energy stored in the Higgs field. How could this energy be released? Once the Higgs field evolved away from the flat part of the energy density curve, it would start to oscillate rapidly about the true-vacuum value. Drawing on the relation between particles and fields implied by quantum field theory, this situation can also be described as a state with a high density of Higgs particles. The Higgs particles would be unstable, however: they would rapidly decay to lighter particles, which would interact with one another and possibly undergo subsequent decays. The system would quickly become a hot gas of elementary particles in thermal equilibrium, just as was assumed in the initial conditions for the standard model. The reheating temperature is calculable and is typically a factor of between two and 10 below the critical temperature of the phase transition. From this point on, the scenario coincides with that of the standard big bang model, and so all the successes of the standard model are retained.

Note that the crucial flaw of the original inflationary model is deftly avoided. Roughly speaking, the isolated bubbles that were discussed in the original model are replaced here by the domains. The domains of the slow-rollover transition would be surrounded by other domains rather than by false vacuum, and they would tend not to be spherical. The term "bubble" is therefore avoided. The key difference is that in the new inflationary model each domain inflates in the course of its formation, producing a vast, essentially homogeneous region within which the observable universe can fit.

Since the reheating temperature is near the critical temperature of the phase transition in the grand unified theory, the matter-antimatter asymmetry could be produced by particle interactions just after the phase transition. The production mechanism is the same as the one predicted by grand unified theories for the standard big bang model. In contrast to the standard model, however, the inflationary model does not allow the possibility of assuming the observed net baryon number of the universe as an initial condition; the subsequent inflation would dilute any initial baryon number density to an imperceptible level. Thus, the viability of the inflationary model depends crucially on the viability of particle theories, such as the grand unified theories, in which baryon number is not conserved.

One can now grasp the solutions to the cosmological problems discussed above. The horizon and flatness problems are resolved by the same mechanisms as in the original inflationary universe model. In the new inflationary scenario the problem of monopoles and domain walls can also be solved. Such defects would form along the boundaries separating domains, but the domains would have been inflated to such an enormous size that the defects would lie far beyond any observable distance. (A few defects might be generated by thermal effects after the transition, but they are expected to be negligible in number.)

Thus, with a few simple ideas the improved inflationary model of the universe leads to a successful resolution of several major problems that plague

NEW INFLATIONARY MODEL deftly evades the horizon, magnetic monopole and domain wall problems. In the two series of drawings on the opposite page, representing the standard big bang model (*left*) and the new inflationary model (*right*), the gray sphere corresponds to the region of space that evolved to become the observed universe and the two-headed green arrow represents the horizon distance. (The relative scales shown here are suggestive only; the actual scales differ by factors that are too extreme to depict.) Three evolutionary stages are shown for each scenario: just before the phase transition (*top*), just after the phase transition (*middle*) and today (*bottom*). In the standard model the horizon distance is always smaller than the gray sphere, making the large-scale uniformity of the observed universe puzzling. Since in the standard model a domain of broken-symmetry phase created in the phase transition would have a radius comparable to the horizon distance, many monopoles and domain walls would be present in the observed universe. In the new inflationary model the horizon distance is always much larger than the gray sphere, and so the observed universe is expected to be uniform on a large scale and to have few, if any, monopoles and domain walls. Just before the phase transition, the gray sphere in the inflationary model is much smaller than it is in the standard model; during the phase transition, the gray sphere in the inflationary model expands by a factor of 10^{50} or more in radius to match the size of the corresponding sphere in the standard model.

the standard big bang picture: the horizon, flatness, magnetic monopole and domain wall problems. Unfortunately the necessary slow-rollover transition requires the fine-tuning of parameters; calculations yield reasonable predictions only if the parameters are assigned values in a narrow range. Most theorists (including both of us) regard such fine-tuning as implausible. The consequences of the scenario are so successful, however, that we are encouraged to go on in the hope that we may discover realistic versions of grand unified theories in which such a slow-rollover transition occurs without fine-tuning.

The successes already discussed offer persuasive evidence in favor of the new inflationary model. Moreover, it was recently discovered that the model may also resolve an additional cosmological problem not even considered at the time the model was developed: the smoothness problem. The generation of density inhomogeneities in the new inflationary universe was addressed in the summer of 1982 at the Nuffield Workshop on the Very Early Universe by a number of theorists, including James M. Bardeen of the University of Washington, Stephen W. Hawking of the University of Cambridge, So-Young Pi of Boston University, Michael S. Turner of the University of Chicago, A. A. Starobinsky of the L. D. Landau Institute of Theoretical Physics in Moscow and the two of us. It was found that the new inflationary model, unlike any previous cosmological model, leads to a definite prediction for the spectrum of inhomogeneities. Basically the process of inflation first smoothes out any primordial inhomogeneities that might have been present in the initial conditions. Then, in the course of the phase transition, inhomogeneities are generated by the quantum fluctuations of the Higgs field in a way that is completely determined by the underlying physics. The inhomogeneities are created on an exceedingly small scale of length, where quantum phenomena are important, and they are then enlarged to an astronomical scale by the process of inflation.

The predicted shape for the spectrum of inhomogeneities is essentially scale invariant, that is, the magnitude of the inhomogeneities is approximately equal on all length scales of astrophysical significance. This prediction is comparatively insensitive to the details of the underlying grand unified theory. It turns out that a spectrum of precisely this shape was proposed in the early 1970s as a phenomenological model for galaxy formation by Edward R. Harrison of the University of Massachusetts at Amherst and Yakov B. Zel'dovich of the Institute of Physical Problems in Moscow, working independently. The details of galaxy formation are complex and are still not well understood, but many cosmologists think a scale-invariant spectrum of inhomogeneities is precisely what is needed to explain how the present structure of galaxies and galactic clusters evolved [see "The Large-Scale Structure of the Universe," by Joseph Silk, Alexander S. Szalay and Yakov B. Zel'dovich; SCIENTIFIC AMERICAN, October 1983].

The new inflationary model also predicts the magnitude of the density inhomogeneities, but the prediction is quite sensitive to the details of the underlying particle theory. Unfortunately, the magnitude that results from the simplest grand unified theory is far too large to be consistent with the observed uniformity of the cosmic microwave background. This inconsistency represents a problem, but it is not yet known whether the simplest grand unified theory is the correct one. In particular, the simplest grand unified theory predicts a lifetime for the proton that appears to be lower than present experimental limits. On the other hand, one can construct more complicated grand unified theories that result in density inhomogeneities of the desired magnitude. Many investigators imagine that with the development of the correct particle theory the new inflationary model will add the resolution of the smoothness problem to its list of successes.

One promising line of research involves a class of quantum field theories with a new kind of symmetry called supersymmetry. Supersymmetry relates the properties of particles with integer angular momentum to those of particles with half-integer angular momentum; it thereby highly constrains the form of the theory. Many theorists think supersymmetry might be necessary to construct a consistent quantum theory of gravity and to eventually unify gravity with the strong, the weak and the electromagnetic forces. A tantalizing property of models incorporating supersymmetry is that many of them give slow-rollover phase transitions without any fine-tuning of parameters. The search is on to find a supersymmetry model that is realistic as far as particle physics is concerned and that also gives rise to inflation and to the correct magnitude for the density inhomogeneities.

In short, the inflationary model of the universe is an economical theory that accounts for many features of the observable universe lacking an explanation in the standard big bang model. The beauty of the inflationary model is that the evolution of the universe becomes almost independent of the details of the initial conditions, about which little if anything is known. It follows, however, that if the inflationary model is correct, it will be difficult for anyone to ever discover observable consequences of the conditions existing before the inflationary phase transition. Similarly, the vast distance scales created by inflation would make it essentially impossible to observe the structure of the universe as a whole. Nevertheless, one can still discuss these issues, and a number of remarkable scenarios seem possible.

The simplest possibility for the very early universe is that it actually began with a big bang, expanded rather uniformly until it cooled to the critical temperature of the phase transition and then proceeded according to the inflationary scenario. Extrapolating the big bang model back to zero time brings the universe to a cosmological singularity, a condition of infinite temperature and density in which the known laws of physics do not apply. The instant of creation remains unexplained. A second possibility is that the universe began (again without explanation) in a random, chaotic state. The matter and temperature distributions would be nonuniform, with some parts expanding and other parts contracting. In this scenario certain small regions that were hot and expanding would undergo inflation, evolving into huge regions easily capable of encompassing the observable universe. Outside these regions there would remain chaos, gradually creeping into the regions that had inflated.

Recently there has been some serious speculation that the actual creation of the universe is describable by physical laws. In this view the universe would originate as a quantum fluctuation, starting from absolutely nothing. The idea was first proposed by Edward P. Tryon of Hunter College of the City University of New York in 1973, and it was put forward again in the context of the inflationary model by Alexander Vilenkin of Tufts University in 1982. In this context, "nothing" might refer to empty space, but Vilenkin uses it to describe a state

devoid of space, time and matter. Quantum fluctuations of the structure of space-time can be discussed only in the context of quantum gravity, and so these ideas must be considered highly speculative until a working theory of quantum gravity is formulated. Nevertheless, it is fascinating to contemplate that physical laws may determine not only the evolution of a given state of the universe but also the initial conditions of the observable universe.

As for the structure of the universe as a whole, the inflationary model allows for several possibilities. (In all cases, the observable universe is a very small fraction of the universe as a whole; the edge of our domain is likely to lie 10^{35} or more light-years away.) The first possibility is that the domains meet one another and fill all space. The domains are then separated by domain walls, and in the interior of each wall is the symmetric phase of the grand unified theory. Protons or neutrons passing through such a wall would decay instantly. Domain walls would tend to straighten with time. After 10^{35} years or more, smaller domains (possibly even our own) would disappear, and larger domains would grow.

Alternatively, some versions of grand unified theories do not allow for the formation of sharp domain walls. In these theories it is possible for different states of broken symmetry states in two neighboring domains to merge smoothly into each other. At the interface of two domains, one would find discontinuities in the density and velocity of matter, and one would also find an occasional magnetic monopole.

A quite different possibility would result if the energy density of the Higgs fields was described by a curve such as the one in the bottom illustration on page 50. As in the other two cases, regions of space would supercool into the false-vacuum state and undergo accelerated expansion. As in the original inflationary model, the false-vacuum state would decay by the mechanism of random bubble formation: quantum fluctuations would cause at least one of the Higgs fields in a small region of space to tunnel through the energy barrier, to the value marked *A* in the illustration. In contrast to the original inflationary scenario, the Higgs field would then evolve very slowly (because of the flatness of the curve near *A*) to its true-vacuum value. The accelerated expansion would continue, and the single bubble would become large enough to encompass the observed universe. If the rate of bubble formation were low, bubble collisions would be rare. The fraction of space filled with bubbles would become closer to 1 as the system evolved, but space would be expanding so fast that the volume remaining in the false-vacuum state would increase with time. Bubble universes would continue to form forever, and there would be no way of knowing how much time had elapsed before our bubble was formed. This picture is much like the old steady state cosmological model on the very large scale, and yet the interior of each bubble would evolve according to the big bang model, improved by inflation.

From a historical point of view, probably the most revolutionary aspect of the inflationary model is the notion that all the matter and energy in the observable universe may have emerged from almost nothing. This claim stands in marked contrast to centuries of scientific tradition in which it was believed that something cannot come from nothing. The tradition, dating back at least as far as the Greek philosopher Parmenides in the fifth century B.C., has manifested itself in modern times in the formulation of a number of conservation laws, which state that certain physical quantities cannot be changed by any physical process. A decade or so ago the list of quantities thought to be conserved included energy, linear momentum, angular momentum, electric charge and baryon number.

Since the observed universe apparently has a huge baryon number and a huge energy, the idea of creation from nothing has seemed totally untenable to all but a few theorists. (The other conservation laws mentioned above present no such problems: the total electric charge and the angular momentum of the observed universe have values consistent with zero, whereas the total linear momentum depends on the velocity of the observer and so cannot be defined in absolute terms.) With the advent of grand unified theories, however, it appears quite plausible that baryon number is not conserved. As a result, only the conservation of energy needs further consideration.

The total energy of any system can be divided into a gravitational part and a nongravitational part. The gravitational part (that is, the energy of the gravitational field itself) is negligible under laboratory conditions, but cosmologically it can be quite important. The nongravitational part is not by itself conserved; in the standard big bang model it decreases drastically as the early universe expands, and the rate of energy loss is proportional to the pressure of the hot gas. During the era of inflation, on the other hand, the region of interest is filled with a false vacuum that has a large negative pressure. In this case, the nongravitational energy increases drastically. Essentially all the nongravitational energy of the universe is created as the false vacuum undergoes its accelerated expansion. This energy is released when the phase transition takes place, and it eventually evolves to become stars, planets, human beings and so forth. Accordingly, the inflationary model offers what is apparently the first plausible scientific explanation for the creation of essentially all the matter and energy in the observable universe.

Under these circumstances, the gravitational part of the energy is somewhat ill defined, but crudely speaking one can say that the gravitational energy is negative and that it precisely cancels the nongravitational energy. The total energy is then zero and is consistent with the evolution of the universe from nothing.

If grand unified theories are correct in their prediction that baryon number is not conserved, there is no known conservation law that prevents the observed universe from evolving out of nothing. The inflationary model of the universe provides a possible mechanism by which the observed universe could have evolved from an infinitesimal region. It then becomes tempting to go one step further and speculate that the entire universe evolved from literally nothing.

FURTHER READING

THE FIRST THREE MINUTES: A MODERN VIEW OF THE ORIGIN OF THE UNIVERSE. Steven Weinberg. Basic Books, Inc. 1977.

INFLATIONARY UNIVERSE: A POSSIBLE SOLUTION TO THE HORIZON AND FLATNESS PROBLEMS. Alan H. Guth in *Physical Review D*, Vol. 23, No. 2, pages 347–356; January 15, 1981.

A NEW INFLATIONARY UNIVERSE SCENARIO: A POSSIBLE SOLUTION OF THE HORIZON, HOMOGENEITY, ISOTROPY AND PRIMORDIAL MONOPOLE PROBLEMS. A. D. Linde in *Physical Letters*, Vol. 108B, No. 6, pages 389–393; February 4, 1982.

COSMOLOGY FOR GRAND UNIFIED THEORIES WITH RADIATIVELY INDUCED SYMMETRY BREAKING. Andreas Albrecht and Paul J. Steinhardt in *Physical Review Letters*, Vol. 48, No. 17, pages 1220–1223; April 1982.

Particle Accelerators Test Cosmological Theory

Is there a limit to the number of families of elementary particles? Debris from the big bang origin of the universe suggests there is, and accelerators are reaching the energies required to confirm the limit

by David N. Schramm and Gary Steigman

Over the past decade two subfields of science, cosmology and elementary particle physics, have become married in a symbiotic relationship that has produced a number of exciting offspring. These offspring are beginning to yield insights on the creation of spacetime and matter at epochs as early as 10^{-43} to 10^{-35} second after the birth of the universe in the primordial explosion known as the big bang. Important clues to the nature of the big bang itself may even come from a theory currently under development, known as the ultimate theory of everything (TOE). A TOE would describe all the interactions among the fundamental particles in a single bold stroke.

The confluence of cosmology and particle physics is even changing the way science is done. Traditionally astronomy has been an observational rather than an experimental science, in which passive observations were made with telescopes and actively controlled experiments have been virtually unknown. Traditionally the tools of particle physics have been high-energy accelerators. Now that cosmology has begun to make predictions about elementary particle physics, it has become conceivable that those cosmological predictions could be checked with carefully controlled accelerator experiments. It has taken more than 10 years for accelerators to reach the point where they can do the appropriate experiments, but the experiments are now in fact in progress. The preliminary results confirm the predictions from cosmology.

It appears therefore that cosmology has become a true science in the sense that ideas not only are developed but also are being tested in the laboratory on time scales that are shorter than a scientist's lifetime. This is a far cry from earlier eras in which cosmological theories proliferated and there was little way to confirm or refute any of them other than on their aesthetic appeal. Conversely, telescopes may eventually be employed to test ideas from fundamental physics, such as proposals for a TOE. Indeed, tests of theories involving interactions of particles with enormous energies could very well have only one available laboratory: the big bang itself.

DAVID N. SCHRAMM and GARY STEIGMAN have been pioneers in developing the interface between cosmology and particle physics. Schramm, who is Louis Block Professor of Physical Sciences at the University of Chicago, received his B.S. in 1967 at the Massachusetts Institute of Technology and his Ph.D. in physics in 1971 from the California Institute of Technology. He enjoys athletics, having climbed mountains throughout the world and written about his experiences for various outdoor magazines. Steigman is professor of physics and astronomy at Ohio State University. He got a B.S. in physics at the City College of the City University of New York in 1961 and a Ph.D. from New York University in 1968. In his spare time he likes to hike, ski, scuba dive and dance.

Among the most exciting offspring produced by the marriage of cosmology and particle physics is the first-begotten. The universe would not look the same if there were too many different fundamental types of elementary particles. In other words, from cosmology it is inferred that the number of fundamental particles must be small. This specific prediction emerged from our analysis of the nuclear reactions that occurred when the universe was roughly one second old. We took the bold step of converting cosmological quantities (such as average energy density) into quantities of interest in particle physics (such as the number of fundamental particles).

Our prediction remains a significant one, but it was particularly powerful at the time it was put forward because the prevailing attitude then was that whenever a particle accelerator attains higher energies, novel particles will be discovered. Theories from particle physics had set no significant limits on how many types of fundamental particles can exist. There appeared to be no end in sight. Now a theoretical prediction made from cosmological considerations contradicted that empirical deduction. As time passes, the prediction has continued to hold up and even looks stronger. The number of elementary particles must be limited, otherwise the universe would be different from the one we know.

To explain how cosmological considerations set limits on the number of types of elementary particles, we must first give a brief overview of particle physics. Over the past half century experiments at particle accelerators have established that fundamental particles can be separated into two broad classes known as fermions and bosons (which are named for the Italian-American physicist Enrico Fermi and the Indian physicist S. N. Bose). Fermions are the particles that make up matter, and bosons are the carriers of the forces between particles. Fermions in turn are divided into two subclasses: quarks and leptons. The word "quark" comes from a curious line in James Joyce's *Finnegans Wake,* "Three quarks for Muster Mark!" and "lepton" comes from the Greek *leptos,* meaning small particle. Quarks are the constituents of neutrons, protons and related particles called hadrons. Leptons, if they carry

Reprinted from the June 1988 issue

an electric charge like that of the electron, can orbit the nucleus to make up atoms; if they are uncharged, like the leptons called neutrinos, they can traverse the entire earth without interacting with anything. Each particle also has an antiparticle with the same mass and lifetime and opposite electrical properties.

The interactions among the various particles are governed by four fundamental forces, each of which is carried by a separate boson or set of bosons. The photon, or quantum of light, carries the electromagnetic force, which couples electric charges; the graviton carries the gravitational force, which couples masses; eight gluons carry the strong nuclear force, which couples quarks; and the intermediate vector bosons carry the weak nuclear force, which is responsible for certain nuclear decays. At present it appears that all interactions in the universe can be reduced to combinations of these four interactions.

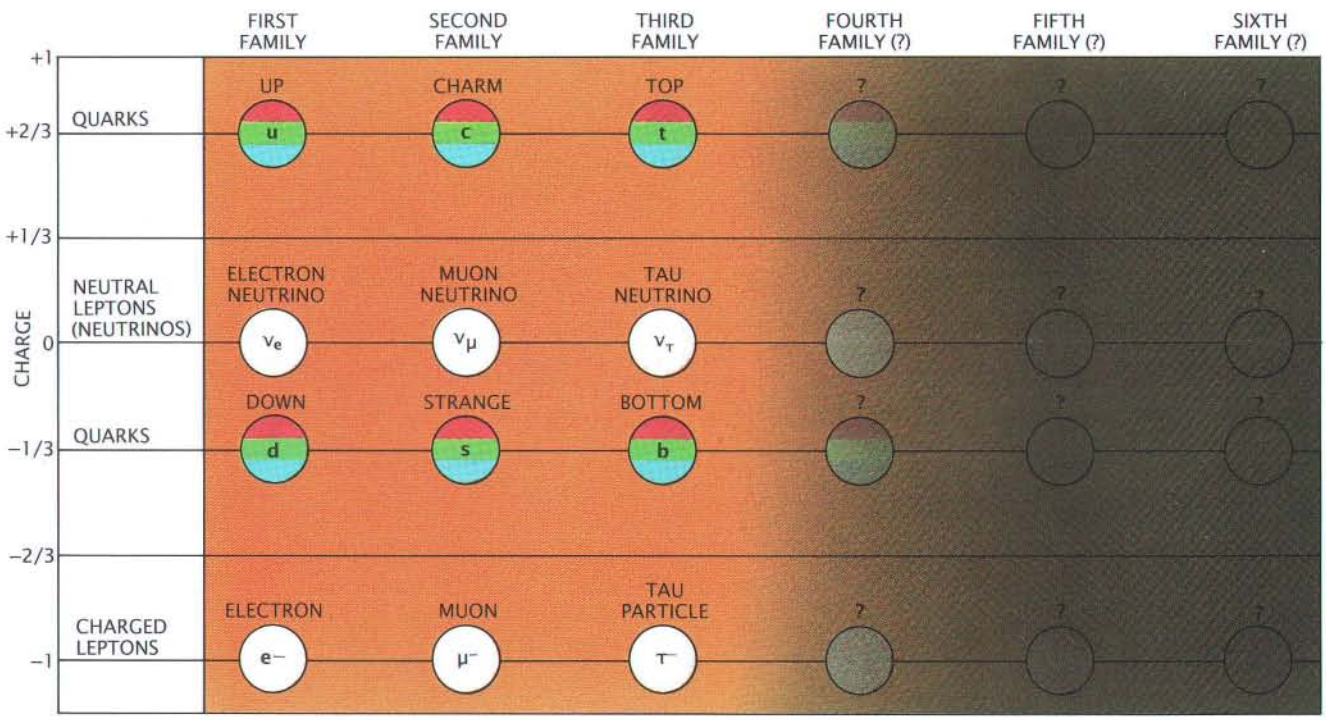
One of the most exciting developments in 20th-century physics has been the proof that at higher energies, or temperatures, the four forces begin to unify. In particular, experiments at CERN, the European laboratory for particle physics, have shown that the weak and the electromagnetic forces merge into a single electroweak force at energies greater than 100 billion electron volts (100 GeV). Such an energy corresponds to the temperature of the universe some 10^{-10} second after the big bang, which was more than four trillion times as great as room temperature. The CERN findings have raised hopes that the strong force will merge with the electroweak force at roughly 10^{15} GeV in some grand unified theory (GUT) and that by 10^{19} GeV the force of gravity will join in to yield a TOE.

The energies needed to test GUT and TOE proposals are enormous compared with the energies accessible to existing particle accelerators. The world's largest accelerator, the Tevatron at Fermi National Accelerator Laboratory, has a circumference of four kilometers and is just reaching energies of two trillion electron volts, or 2,000 GeV. An accelerator similar to the Tevatron but scaled up to energies at which GUT proposals could be tested would stretch to the nearest stars, and a TOE machine would reach to the galactic center. Both machines are beyond even the most optimistic science budgets. This simple fact has been one of the driving forces behind the attempts to utilize cosmological observations to test predictions of particle physics.

The flow of information has also begun to go the other way: the accelerators of particle physics are being employed to test a prediction of cosmology. The cosmological prediction we have been concerned with pertains to setting limits on the number of fundamental particles of matter.

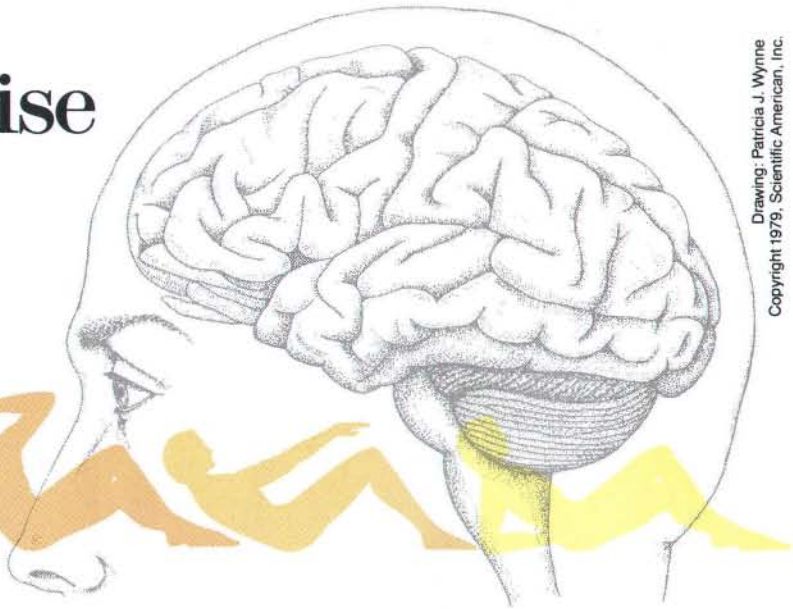
It appears that there are 12 fundamental particles, as well as their corresponding antiparticles. Six of the fundamental particles are quarks, and they carry the whimsical names of "up," "down," "charm," "strange," "top" (or "truth") and "bottom" (or "beauty"). All the quarks have been discovered except the top quark; although theoretical arguments for the existence of that quark are strong, experimental evidence is at the moment absent. The other six fundamental particles are leptons: the electron, the muon, the tau particle, and three neutrinos associated with each of them (namely, the electron neutrino, the muon neutrino and the tau neutrino).

The 12 particles are grouped in three families, each family consisting of four members. The first family is made up of the up and down quarks, the electron and the electron neutrino; the second family consists of the charm and strange quarks, the muon and the muon neutrino, and the third family of



FUNDAMENTAL CONSTITUENTS OF MATTER, called quarks and leptons, are grouped into families consisting of two types of each class of particles. The particles can be distinguished by their electric charge, among other properties. At present there are three families made up of 12 known quarks and leptons. All ordinary matter is composed of members of the first family. (The proton, for instance, consists of two "up" quarks and one "down" quark.) Theories offer few predictions about how many families should exist; in principle, the number could be infinite. Cosmological theories, however, suggest an upper limit of four families. A test of this limit is now being carried out at several particle accelerators.

Not all exercise is physical.



Drawing: Patricia J. Wynne
Copyright 1979, Scientific American, Inc.

If order card is missing,
call Toll Free 1-800-333-1199.

In fractions of a second, your mind can have you slamming a ball off two walls, or sliding into second base.

It decides, directs and transmits the necessary brain waves needed to enjoy an hour of hard, body-toning physical exercise. It's what intelligent people do to keep themselves fit and healthy.

It's a well known fact that we use only a small portion of our minds' capabilities and like our bodies, our minds thrive on good healthy use. SCIENTIFIC AMERICAN provides an ideal workout for the mind, stretching the imagination and toning the intellect.

SCIENTIFIC AMERICAN probes the opposing mysteries of outer and inner space. Rocket technology, genetic manipulation, the socio-economic effect of earthquakes on populations or the origins of Indo-European languages—just a small sampling of the kinds of articles written by scientists and edited for your active mind.

Each month, SCIENTIFIC AMERICAN brings the individual talents of scientists, editor and artist together to keep you up-to-date on the new developments and important technologies that affect our everyday lives.

Each article is an opportunity to share in the act of science and observe first hand the arrival at discovery.

Join us at the frontiers of knowledge. Subscribe to

SCIENTIFIC AMERICAN

and exercise your mind.

12 issues per year $19.97

Foreign subscriptions are $38.00
(new subscribers only)

SAVE OVER $27.00 off the cover price.

Simply complete and mail the attached card to enter your subscription today!

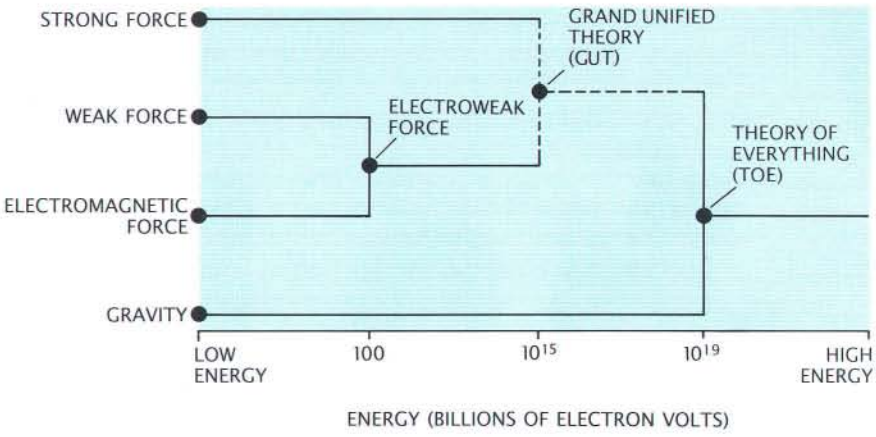
the top and bottom quarks, the tau particle and the tau neutrino. All ordinary matter is made of members of the first family. The proton, for instance, consists of two up quarks, each bearing two thirds of a unit of positive electric charge, and a down quark, which carries a charge of one third of a unit of negative electric charge. The neutron consists of two down quarks and an up quark. Every atom is simply a hard kernel of tightly bound protons and neutrons surrounded by a cloud of electrons.

Since all ordinary matter is made of members of the first family, one of the great mysteries of particle physics today is why the other families exist and how many of them there are. As the late I. I. Rabi said when the muon was discovered, "Who ordered that?" Based on the trend of more particles being found as the energy of particle accelerators is increased, one might expect families to continue to proliferate. Actually GUT proposals have said little about the total number of families. For example, the first GUT model to gain popularity in the mid- to late 1970s, SU(5) (for special unitary group with five degrees of freedom), can have any number of families.

There is a good reason for having at least three families, however. M. Kobayashi of the Japanese proton accelerator, KEK, and T. Masakawa of the University of Tokyo have pointed out that the asymmetry between matter and antimatter observed in 1964 by Val L. Fitch of Princeton University and James W. Cronin of the University of Chicago is best understood if there are at least three families of elementary particles. The asymmetry may provide the explanation for the observed excess of matter over antimatter in the universe, thereby enabling matter to exist [see "A Flaw in a Universal Mirror," by Robert K. Adair; SCIENTIFIC AMERICAN, February 1988]. Having at least three families is not quite useless.

Yet one would still like to know precisely how many families of quarks and leptons there are. If quarks and leptons are the fundamental building blocks of nature, one would like to be acquainted with all the components. If it appeared that the number of families were unlimited, one would question whether quarks and leptons are truly fundamental. Just as atoms are made up of protons, neutrons and electrons, so perhaps quarks and leptons are made up of still smaller entities [see "The Structure of Quarks and Leptons," by Haim Harari; SCIENTIFIC AMERICAN, April 1983].

It is now clear that an answer to the question of the number of families of quarks and leptons could well come from cosmology. Cosmology suggests there must be a finite number of families and, further, limits the possible range to small values: only three or at most four families exist.



FOUR FORCES account for all known interactions among elementary particles. The strong force couples quarks, the weak force is responsible for certain nuclear decays, the electromagnetic force couples electric charge and gravity couples masses. It is thought the four forces were once unified at the higher energies characteristic of the universe soon after the big bang, and indeed a theory that unifies the weak force and the electromagnetic force has already been verified. A grand unified theory (GUT) would unify these forces with the strong force; a theory of everything (TOE) would describe all four forces as aspects of a single force.

The prediction of the limit to the number of families is based on evidence garnered by observing the debris from the greatest accelerator experiment of all, the big bang. The big bang model of the universe began as one of two rival theories of cosmology dominating discussion in the 1950s and early 1960s. The other theory was called the steady state theory. Both were developed to account for Edwin P. Hubble's discovery in 1929 that the universe is expanding. The big bang model holds that the universe was once hot and enormously dense and that as it has expanded it has cooled and become less dense. The steady state theory holds that matter is continuously created, so that as the universe has expanded, its density has remained constant.

In the 1960s the big bang model received several observational boosts, and by the 1970s it was the clear winner. The most publicized of these boosts was the Nobel Prize–winning discovery made by Arno A. Penzias and Robert W. Wilson of Bell Telephone Laboratories. If the big bang model is correct, at one time the universe would have been sufficiently dense and so hot that the matter in the universe would have generated a characteristic spectrum of thermal radiation. According to the steady state theory, on the other hand, the density of the universe has always been what it is today, and so the universe never existed in a dense, hot state. Hence, there should be no thermal radiation. Penzias and Wilson discovered a microwave background radiation that is consistent with the hot, dense scenario expected with the big bang.

The strongest support for the big bang model comes from studies of primordial nucleosynthesis: the formation of the elements. Temperatures nearly 100 million times as great as room temperature are needed to forge many elements from protons and neutrons; such temperatures would have occurred about one second after the big bang. By measuring the relative abundances of elements, therefore, one can probe conditions as far back as one second after creation. In comparison, the microwave background radiation serves as a probe of the universe back only to 100,000 years after creation, when photons last scattered with matter at temperatures of some 3,000 kelvins (degrees Celsius above absolute zero), or approximately 10 times room temperature.

We shall delve into the details of big bang nucleosynthesis below, since they not only help to establish the big bang but also lead to the connection with particle physics. But it is first worth noting that the theory of big bang nucleosynthesis has predicted the abundances of several light elements and their isotopes, including helium 3, helium 4, deuterium (the heavy isotope of hydrogen) and lithium 7. The predicted

abundances span almost 10 orders of magnitude. Observations appear to verify all of these predictions in quantitative detail.

The impressive agreement between the theoretical predictions of big bang nucleosynthesis and the astronomically observed abundances of the light elements provides a bonus. Agreement between theory and observation occurs for a value of the density of protons and neutrons that is completely consistent with the density determined from studies of the dynamics of luminous matter in the universe. The predictions based on the evolution of the universe during the first 1,000 seconds after the big bang are consistent with observations made some 10 billion years later.

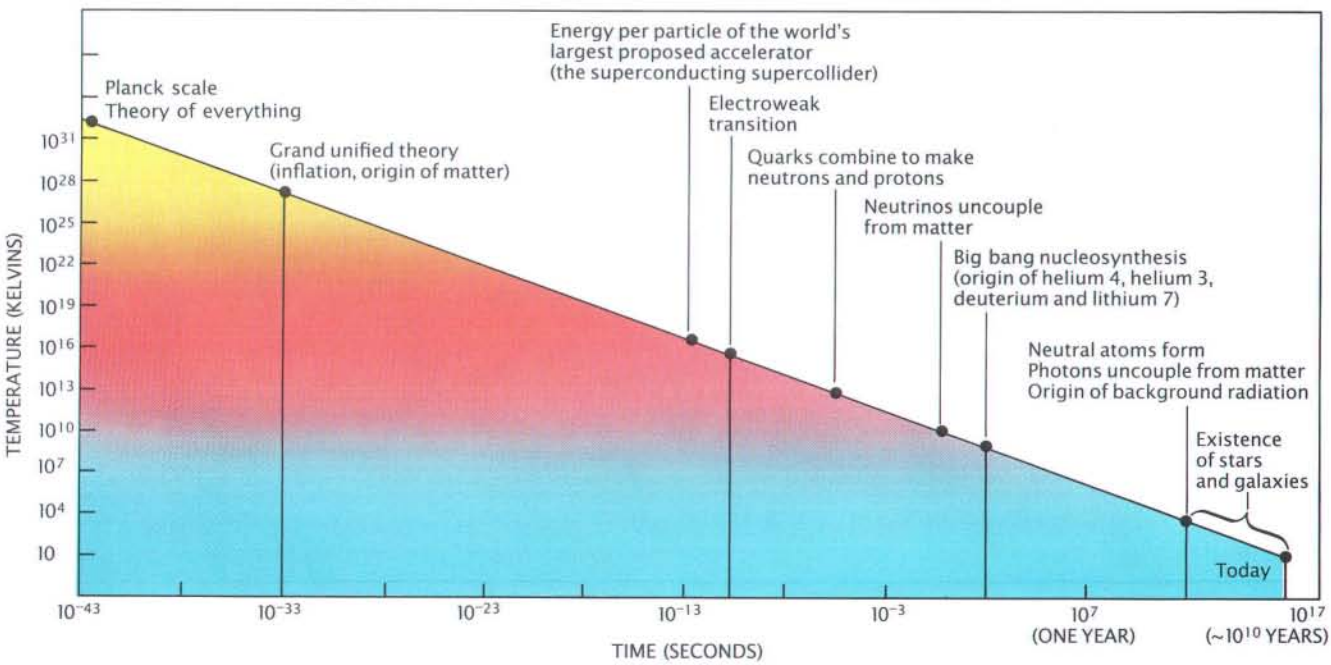
Physicists now seem to have a *quantitative* understanding of the behavior of the universe back to the time of big bang nucleosynthesis. This detailed understanding has provided the confidence needed in attempting to push back to even earlier times appropriate for a GUT or a TOE.

The power of the theory of big bang nucleosynthesis derives from the fact that essentially all the input into the relevant equations is well known from laboratory experiments. In particular, the temperatures at which big bang nucleosynthesis is thought to have occurred correspond to energies that are easily explored with relatively low-energy accelerators such as Van de Graaff generators. As a consequence, the behavior of atomic nuclei under the conditions of big bang nucleosynthesis is not a matter of guesswork; it is precisely known.

To calculate what happens, all one has to do is follow the evolution of a gas of neutrons and protons in an expanding and cooling universe. Since neutrons and protons are collectively referred to as nucleons, physicists call such a gas a nucleon gas. At temperatures much greater than 10 billion kelvins, which correspond to times well before the first second after the big bang, the protons and neutrons were in equilibrium and were present in equal numbers. These temperatures were too hot to allow protons and neutrons to fuse into more complex nuclei. Collisions with electrons and positrons (antielectrons) and neutrinos and antineutrinos changed neutrons into protons and protons into neutrons at approximately equal rates. The neutron is slightly heavier than the proton, and so neutrons change into protons more easily than protons change into neutrons. When the energies were very high, however, the mass difference had a negligible effect.

As the temperature of the universe fell to 10 billion kelvins, the mass difference became more significant, and the ratio of neutrons to protons dropped from one to less than a third. By the time the universe reached a billion kelvins, the ratio was slightly below a seventh. At that point the temperature was cool enough to allow protons and neutrons to begin to fuse into the simplest complex nucleus: deuterium, which consists of a single proton and neutron. Interactions of deuterium with other protons and neutrons produced tritium (a proton and two neutrons) and helium 3 (two protons and a neutron). These nuclei in turn interacted to produce helium 4 (two protons and two neutrons). Since helium 4 is much more tightly bound than any other light nucleus, the flow of reactions converted almost all the neutrons that existed at a billion kelvins into helium 4. A small amount of beryllium 7 (four protons and three neutrons) and of lithium 7 (three protons and four neutrons) was produced when helium 4 interacted with helium 3 and tritium, respectively. In short, the big bang nucleosynthesis is believed to have generated helium 4 with traces of deuterium, helium 3 and lithium 7.

The flow essentially ceased at helium 4 because no stable nuclei are produced when a helium 4 nucleus interacts with a proton, a neutron or another helium 4 nucleus. The majority of the other elements were produced inside stars, which have densities suffi-



THERMAL HISTORY OF THE UNIVERSE, starting 10^{-43} second after the big bang and continuing to the present, shows that most of the helium 4, helium 3, deuterium (heavy hydrogen) and lithium 7 in the universe was synthesized about a minute after the big bang. Heavier elements were forged tens of millions to billions of years later in the interior of stars.

cient to allow three helium 4 nuclei to combine to make carbon 12.

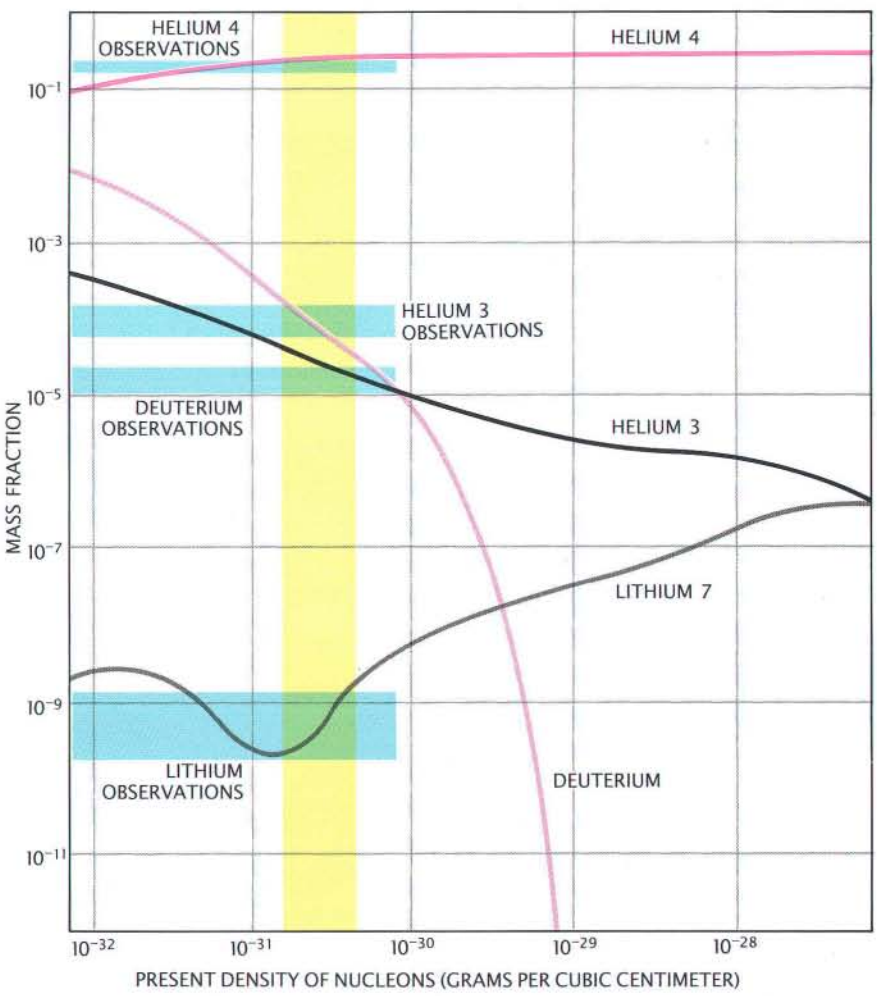
The abundances of the light elements predicted by the theory of big bang nucleosynthesis, as we have mentioned, agree quite nicely with the observed abundances. According to the theory of big bang nucleosynthesis, matter began to coalesce when the ratio of neutrons to protons was a seventh. Because virtually all neutrons were swept up into helium 4 nuclei (which contain equal numbers of protons and neutrons), the abundance of helium 4 should account for about a fourth of the total mass of ordinary matter. Actually the observed abundance of helium in galaxies, including our own, ranges from about 20 to 30 percent. The predicted abundances of deuterium, helium 3 and lithium 7, which range from less than one part in 10,000 to as little as one part in 10 billion, also match the abundances that are observed.

How does the theory of big bang nucleosynthesis set limits on the allowed number of families of elementary particles? Quite simply, if the number of particle families exceeded three or four, the predicted abundance of helium 4 would exceed the observed abundance.

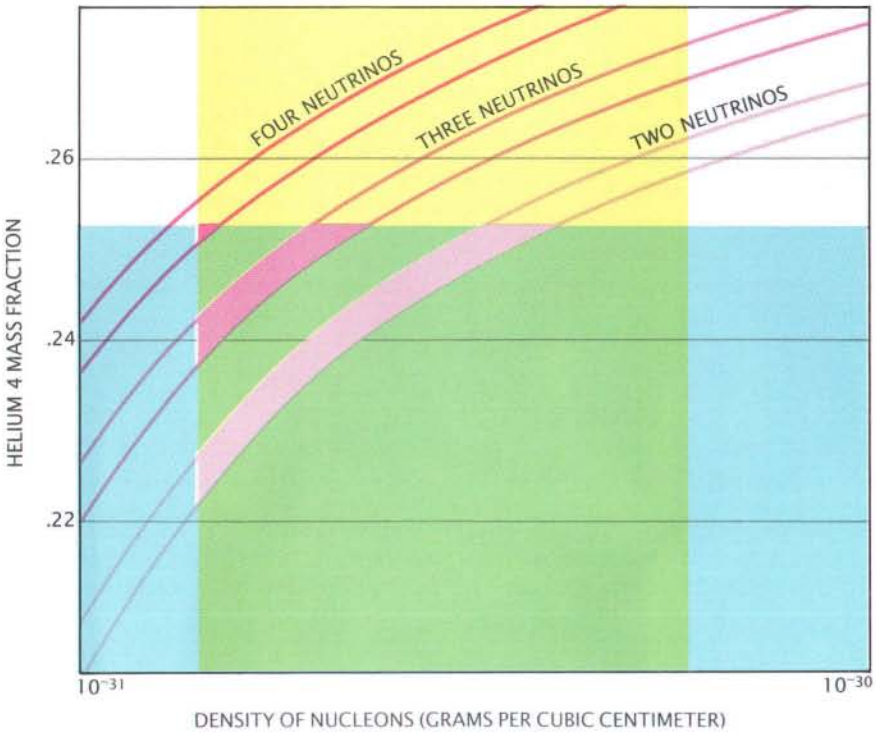
The reason such a statement can be made is that the predicted abundances of the light elements depend on only two variables: the density of nucleons and the density of radiation in the universe. Unfortunately, the value of neither of these variables is precisely known. It turns out, however, that only a small range of values for each of the two variables produces abundances that are consistent with the observed abundances. By substituting the values of the observed abundances into the appropriate equations, one can determine what the density of nucleons and of radiation must be. Knowing those values leads to a number of interesting conclusions.

The density of a nucleon gas increases in proportion to the cube of the temperature, so that when the universe was twice as hot as it is today, it was eight times as dense in nucleons. By determining what the nucleon density must have been during nucleosynthesis to produce the abundances of deuterium, helium 3 and lithium 7 seen today, one can calculate the present nucleon density. It is between 2×10^{-31} and 5×10^{-31} gram per cubic centimeter. Such a range of values is consistent with the density of luminous material inferred from studies of the dynamics of galaxies and clusters of galaxies, but it is at least 10 times less than the estimated density of gravitational mass needed to close the universe, or halt the big bang expansion. If the universe is closed, additional nonnucleonic matter is needed. The search for such matter, which would be dark, or invisible to telescopes, and made up of something other than nucleons, is now under way [see "Dark Matter in the Universe," by Lawrence M. Krauss; SCIENTIFIC AMERICAN, December 1986].

An analogous line of reasoning for the radiation density provides the constraints on the number of families of elementary particles. The radiation density is important for big bang nucleosynthesis because it controlled the rate of expansion of the universe during that time. The radiation density at any time is proportional to the number of types of radiation or, equivalently, the number of types of particles moving at nearly the speed of light. During big bang nucleosynthesis, it is thought there were nine types of relativistic particles: the photon (of course), the electron and the positron, the electron neutrino, the muon neutrino and the tau neutrino and their three antiparticles. Neutrinos are either massless or have such a small mass that they travel at nearly the speed of light; the electron and the positron have a mass that is small enough so that at the high energies achieved at the time of primordial nucleosynthesis, they too would have traveled at close to the speed of light. The radiation density associated with the nine kinds of particles leads to conditions that would have been just right to produce the observed abundance of helium 4. (Interestingly, the helium 4 density is virtually indepen-



PREDICTED ABUNDANCES of helium 4, helium 3, deuterium and lithium 7 in the big bang model of the universe (*curves*) closely agree with the observed abundances (*horizontal bands*). The predicted abundances change as a function of the density of nucleons (protons and neutrons) at the time of the big bang; the vertical band indicates the best cosmological estimate of that density today. The close agreement is one of the strongest arguments supporting the big bang model.



HELIUM 4 ABUNDANCE suggests there are at most four families of elementary particles. The three curves represent an enlargement of the part of the helium 4 curve lying within the shaded vertical band in the illustration on the preceding page; the narrow curve in that illustration resolves into the three broad curves. The bottom curve shows the helium 4 abundance predicted if there were two families of particles. The middle curve shows the abundance predicted for three families, and the top curve shows the abundance predicted if there are four families. The predicted abundances of helium 4 for two and three families of particles are well within the region defined by helium 4 observations and estimates of nucleon density (*green region*). A fourth family would produce an abundance that would be very close to the allowed extremes. Clearly, there is no room for any more than four families.

dent of the nucleon density, a fact that Fred Hoyle and his colleague Roger Taylor first noted in the 1960s and that was later verified more rigorously by a number of investigators.)

In calculating the abundance of helium 4, we have allowed for photons, electrons and the three known neutrino species, as well as their antiparticles. If other families of fundamental particles exist, the calculation would have to be modified. The only new family members that would affect the calculation are neutrinos, since in any family beyond the first, they alone are light enough to travel close to (or at) the speed of light. Presumably each new family beyond the third would contribute one neutrino and a corresponding antineutrino.

If the gas from which the universe was made had contained additional neutrinos, its density of radiation would have been greater. As a consequence, the cosmological expansion during the period of big bang nucleosynthesis would have been more rapid. It so happens that the ratio of neutrons to protons is quite sensitive to the rate of cosmological expansion. A higher rate of expansion would have meant that neutrons would have had less time to change into protons, and so more neutrons would have been left: the ratio of neutrons to protons would have been greater. Because neutrons were quickly swept up into helium 4 nuclei, the abundance of helium 4 would be greater.

Helium 4 is the most abundant of the nuclei synthesized in the big bang, and so it is the element observers can measure with the most accuracy. Since helium 4 is produced in stars, however, it is important to estimate what part of the helium observed in astronomical objects is primordial—from the big bang—and what part was generated by stars after the big bang. In collaboration with John S. Gallagher of the Lowell Observatory, we have found that the additional amount of helium 4 produced by stars can be tracked by measuring the carbon content of objects; stars that make helium also make carbon, so that the helium abundance increases with the carbon abundance. This allows one to "subtract" the contribution of stars to the helium 4 abundance from the observed abundance in order to infer the true primordial abundance. We have determined that the highest allowed value of the primordial abundance of helium 4 is slightly less than 25 percent.

Our calculations show that such a low abundance could have emerged from the big bang only if there is no more than one additional kind of neutrino and corresponding antineutrino. If more neutrinos had existed, the radiation density would have been so great that the amount of helium 4 produced during the big bang nucleosynthesis would be greater than the observed abundance. In other words, the total number of families of elementary particles is three or at most four. Our finding suggests that all the fundamental families of elementary particles may already have been discovered. This basic argument was made by us more than 10 years ago in collaboration with James E. Gunn of Princeton University; subsequently, the measurements of the helium 4 abundance and the estimate of its primordial value have improved significantly. What makes the argument particularly exciting and timely today is that accelerators are now beginning to check it.

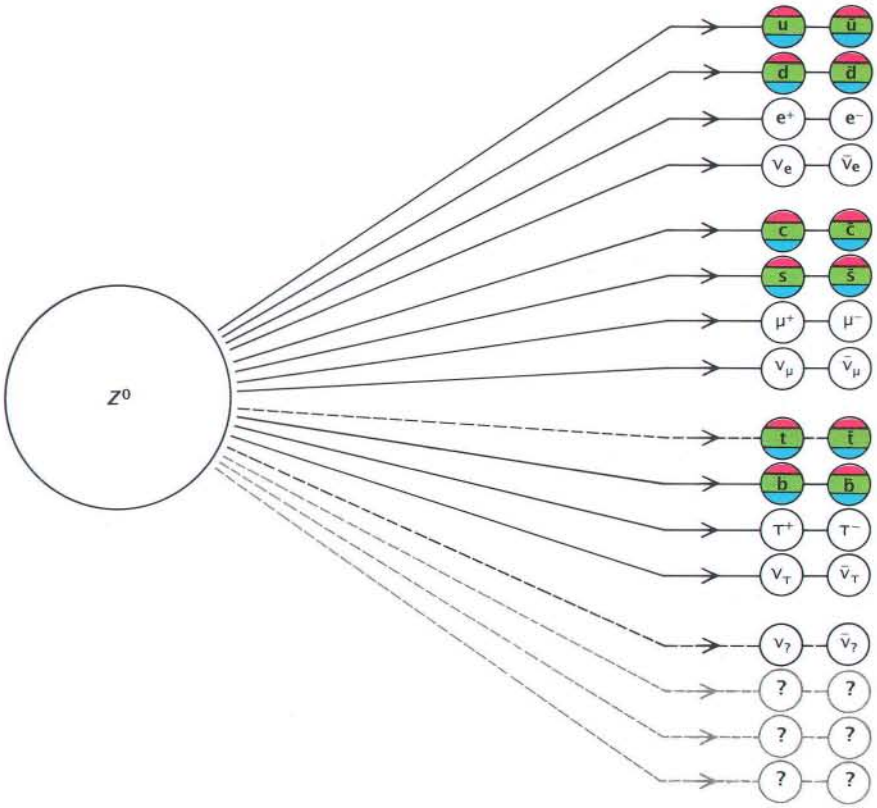
Searching for new kinds of neutrinos has always been a difficult and tedious procedure. Traditionally the only means of discovering a neutrino has been to produce its associated charged lepton first. The drawback of the approach is that even though neutrinos are quite light or even massless, a great deal of energy is necessary to generate the associated leptons; the heavier the mass of the associated lepton is, the greater the energy of the accelerator must be to generate it. The tau particle has such a high mass, for instance, that Martin L. Perl and his collaborators at the Stanford Linear Accelerator Center (SLAC) needed several billion electron volts of energy, corresponding to temperatures exceeding 10^{13} kelvins, to find it. With such an approach, one can always argue that the next lepton is just beyond the limits of the current accelerators.

The new way of searching for neutrinos grew out of the CERN experiments mentioned above, which have shown that the weak and the electromagnetic forces are actually aspects of a single electroweak force. In 1983 the CERN investigators, a team of hundreds of physicists led by Carlo

Rubbia, successfully accomplished what they set out to do: they proved the existence of the intermediate vector bosons, the conjectured carriers of the weak nuclear force. Three such particles were found, the W^+, the W^- and the Z^0 bosons.

The discovery of the Z^0 boson was particularly relevant to our work. The Z^0 boson is electrically neutral, which means that it can decay to pairs of neutrinos and antineutrinos, since they are also electrically neutral. (The Z^0 boson can also decay to charged particle-antiparticle pairs such as electrons and positrons.) In other words, the Z^0 boson allows one to produce each kind of neutrino directly, without first producing the associated lepton. The lifetime of the Z^0 boson serves as a measure of the number of families of elementary particles, because the more families there are, the more options the particle has for its decay. Hence, a greater number of families should mean a shorter lifetime for the Z^0 boson. Careful measurements of the lifetime of the Z^0 boson could therefore reveal the number of families of elementary particles.

To measure the decay properties of the Z^0 boson, one must first have a machine with enough energy to produce that particle. Older accelerators, in which high-energy beams of protons or electrons struck stationary targets, spent most of their energy on motion, leaving relatively little energy available for producing particles. The novel approach taken with the CERN machine, utilizing an idea by Simon van der Meer, is to have protons and antiprotons collide head-on so that most of the energy can be utilized in producing new particles.



TESTS OF FAMILY NUMBER are under way at various particle accelerators around the world. A particle called the Z^0 boson can decay into quarks and leptons, each particle paired with its antiparticle. (An antiparticle has the same mass as its corresponding particle but opposite electrical properties; it is often represented by the symbol for the corresponding particle with a horizontal bar over it.) The allowed decays are represented by solid lines. The more families there are, the more decay routes the Z^0 boson should have and therefore the shorter its lifetime should be. As a result, measurements of the lifetime should indicate the number of families. Current estimates place the limit at five families of particles. Experiments in the future should result in more precise results.

Several accelerators at sites around the world now employ head-on collisions. The Tevatron at Fermilab collides protons with antiprotons; SLAC and the Deutsches Elektronen-Synchrotron (DESY) collide electrons with positrons. Although the energies of the last two are too low to produce real Z^0 bosons, through quantum-mechanical phenomena they can sometimes produce "virtual" particles that mimic the effects of the Z^0 boson.

Preliminary results from the machines indicate that there are at most five families of elementary particles. David B. Cline of the University of California at Los Angeles and the University of Wisconsin at Madison and a leader of the CERN neutrino-counting efforts has shown that the lifetime of the Z^0 boson is about what one would expect with three families. Experimental uncertainties, however, allow for two additional kinds of neutrinos and hence two additional families. Theodore L. Lavine, a graduate student at Wisconsin, has combined data from SLAC and DESY and obtained a comparable limit on the total number of neutrinos of about five. For the first time, accelerators are counting neutrino types and getting a small number, one that was predicted by cosmological theory, not by particle theory.

The next step promises to be even more exciting. As new accelerators are completed and begin producing more data with fewer uncertainties, the cosmological limit of three or at most four families will be checked with extreme accuracy. The SLAC machine is being modified to generate copious numbers of Z^0 bosons; the new accelerator is called the Stanford Linear Collider (SLC). Another accelerator under construction at CERN called the Large Electron-Positron (LEP) collider ring will also produce large numbers of Z^0 bosons. The machines will probe the early universe with an effectiveness that no telescope will ever match.

FURTHER READING

COSMOLOGICAL LIMITS TO THE NUMBER OF MASSIVE LEPTONS. Gary Steigman, David N. Schramm and James E. Gunn in *Physics Letters,* Vol. 66B, No. 2, pages 202–204; January 17, 1977.

THE EARLY UNIVERSE AND HIGH-ENERGY PHYSICS. David N. Schramm in *Physics Today,* Vol. 36, No. 4, pages 27–33; April 1983.

BIG BANG NUCLEOSYNTHESIS: THEORIES AND OBSERVATIONS. Gary Steigman and Ann Merchant Boesgaard in *Annual Review of Astronomy and Astrophysics,* Vol. 23, pages 319–378; 1985.

NEUTRINO FAMILIES: THE EARLY UNIVERSE MEETS ELEMENTARY PARTICLE/ACCELERATOR PHYSICS. David B. Cline, David N. Schramm and Gary Steigman in *Comments on Nuclear and Particle Science,* Vol. 17, No. 3, pages 145–161; 1987.

BroadBa

Or, Why Broadband Services From Your Local Telco Will Be Music To Your Network.

Broadband services unleash the true power of the public switched network. Let networks of computers listen and talk. Allow you to bring up a remotely stored document in one window. Video conferences in two others. How? By using fast packet switching, it allows existing public networks to transmit more information at faster speeds. So things impossible now, will be possible tomorrow with broadband. High-definition television. Interactive education. Image processing. High-resolution faxes. To learn exactly what broadband services can do for you and your business, talk to the broadband leaders. Call your local phone company or AT&T Network Systems at 1 800 638-7978, ext. 6110.

AT&T and Your Local Phone Company Technologies For The Real World.

AT&T
Network Systems

The Mystery of the Cosmological Constant

According to theory, the constant, which measures the energy of the vacuum, should be much greater than it is. An understanding of the disagreement could revolutionize fundamental physics

by Larry Abbott

What determines the structure of space and time in the universe? According to Einstein's general theory of relativity, the geometric properties of space are related to the density of energy (and momentum) in the universe. To understand the structure of space-time, therefore, we must identify potentially relevant sources of energy and evaluate their contributions to the total energy (and momentum) density. The most obvious energy sources that come to mind are ordinary matter and radiation. A much less obvious source of energy that can have an enormous impact on the structure of the universe is empty space itself: the vacuum.

The notion that the vacuum can be a source of energy may at first seem counterintuitive. But present theories of elementary particles and forces not only allow for a nonzero vacuum energy density but also strongly suggest that it should have a large value. Is the vacuum energy density really as large as these theories appear to suggest it is?

The answer is most emphatically no. The geometric structure of the universe is extremely sensitive to the value of the vacuum energy density. So important is this value that a constant proportional to the vacuum energy density has been defined. It is called the cosmological constant. If the vacuum energy density, or equivalently the cosmological constant, were as large as theories of elementary particles suggest, the universe in which we live would be dramatically different, with properties we would find both bizarre and unsettling. What has gone wrong with our theories? We do not know the answer to this question at present. Indeed, a comparison of our theoretical and experimental understanding of the cosmological constant leads to one of the most intriguing and frustrating mysteries in particle physics and relativity today.

Most people are unaccustomed to the idea that the vacuum might have a nonzero energy density: How can a unit volume of empty space contain energy? The answer in part lies in the fact that, according to quantum mechanics, physical quantities tend to fluctuate unavoidably. Even in the apparent quiet of the vacuum state, pairs of particles are constantly appearing and disappearing. Such fluctuations contribute energy to the vacuum.

The notion of a vacuum energy is also unfamiliar because that energy cannot be detected by normal techniques. Energies are usually determined by measuring the change in the energy of a system when it is modified in some way or by measuring a difference in energy between two systems. For example, we might measure the energy released when two chemicals react. Because of this, energy as we normally define it is a relative quantity. The energy of any state of a system only has meaning in relation to some other state.

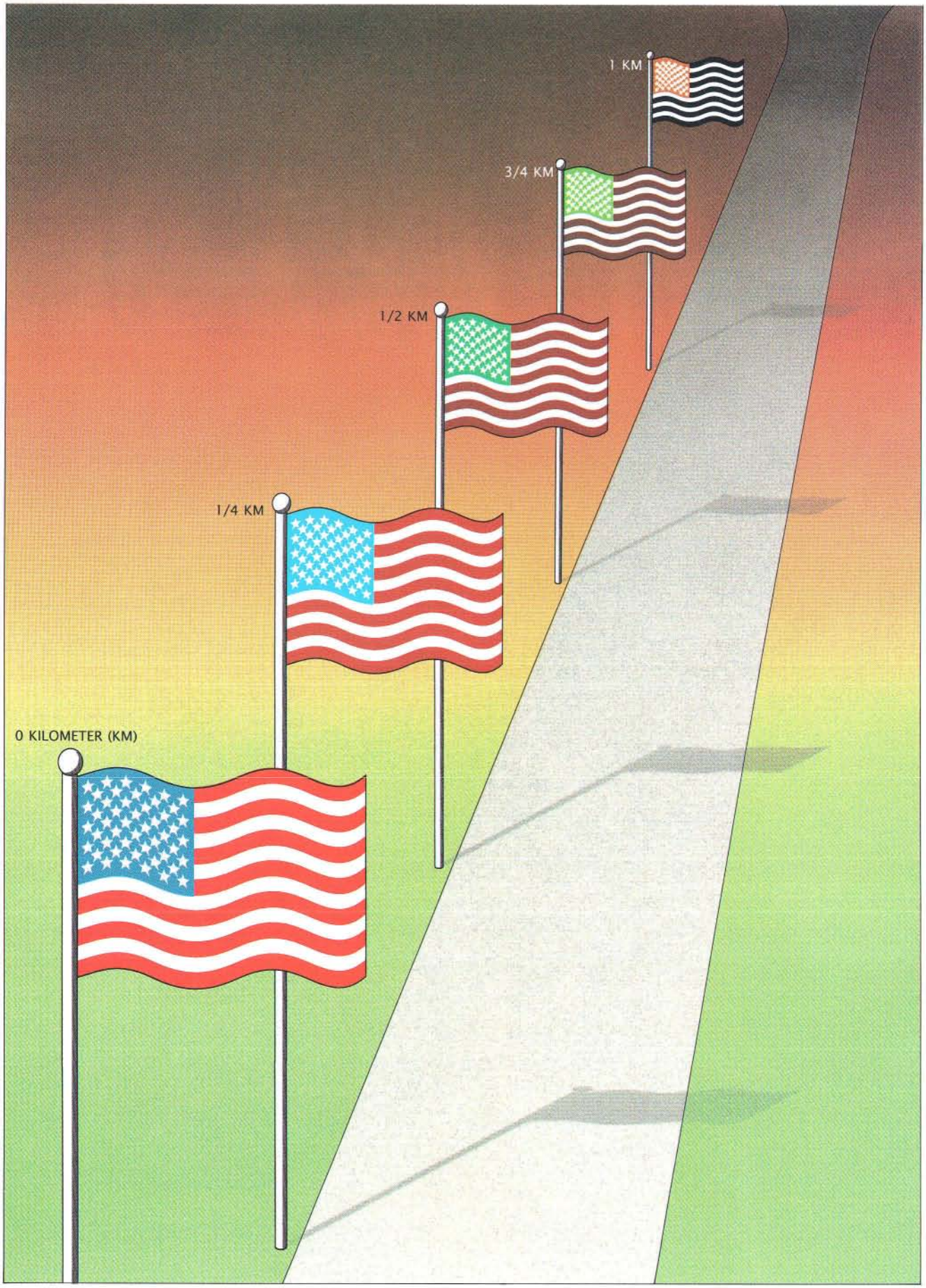
By convention, energies are often measured in relation to the vacuum. When it is defined in this way, the vacuum automatically has zero energy in relation to itself. The traditional approach will not work if we want to discuss the energy of the vacuum in an absolute and significant way. We must use a different technique to measure its value.

The only way to establish an absolute measure of energy is by using gravity. In general relativity, energy is the source of gravitational fields in the same way that electric charge is the source of electric fields in the Maxwell theory of electromagnetism. An energy density of any kind, including that produced by fluctuations in the vacuum, generates a gravitational field that reveals itself as a change in the geometry of space-time.

LARRY ABBOTT is professor of physics at Brandeis University. He earned his Ph.D. in physics from Brandeis in 1977. He joined the faculty there in 1979 after working at the Stanford Linear Accelerator Center and at CERN, the European laboratory for particle physics near Geneva. Abbott writes: "Over the past several years I have been interested in the application of new ideas in particle physics to cosmology and have worked on inflationary cosmology, dark matter and the large-scale structure of the universe. At the same time, I have become intrigued by the problem of the cosmological constant. More recently I have studied the implications of the observation of neutrinos from supernova 1987A and have become interested in the physics of neural networks."

UNIVERSE with a large cosmological constant would be vastly different from the existing one. Here an artist has painted a scene as it might appear if the constant were as large as theoretical estimates suggest it could be. The illustration is based on a positive value for the constant on the order of $1/(1\ \text{kilometer})^2$. With such a value the structure of space would be so distorted that the radiation from distant objects would be redshifted, or shifted toward longer wavelengths. The farther an object is from an observer, the greater the redshift would be. A spectral blue object about a kilometer away would look red; objects more than a kilometer or so away would have such large redshifts that they would be invisible. Distant objects would appear spatially distorted.

Reprinted from the May 1988 issue


1 KM
3/4 KM
1/2 KM
1/4 KM
0 KILOMETER (KM)

COSMOLOGICAL CONSTANT = $8\pi G/c^4 \times$ VACUUM ENERGY DENSITY

Here G is Newton's gravitational constant and c is the speed of light. Defined in such a way, the cosmological constant has units of 1 over distance squared.

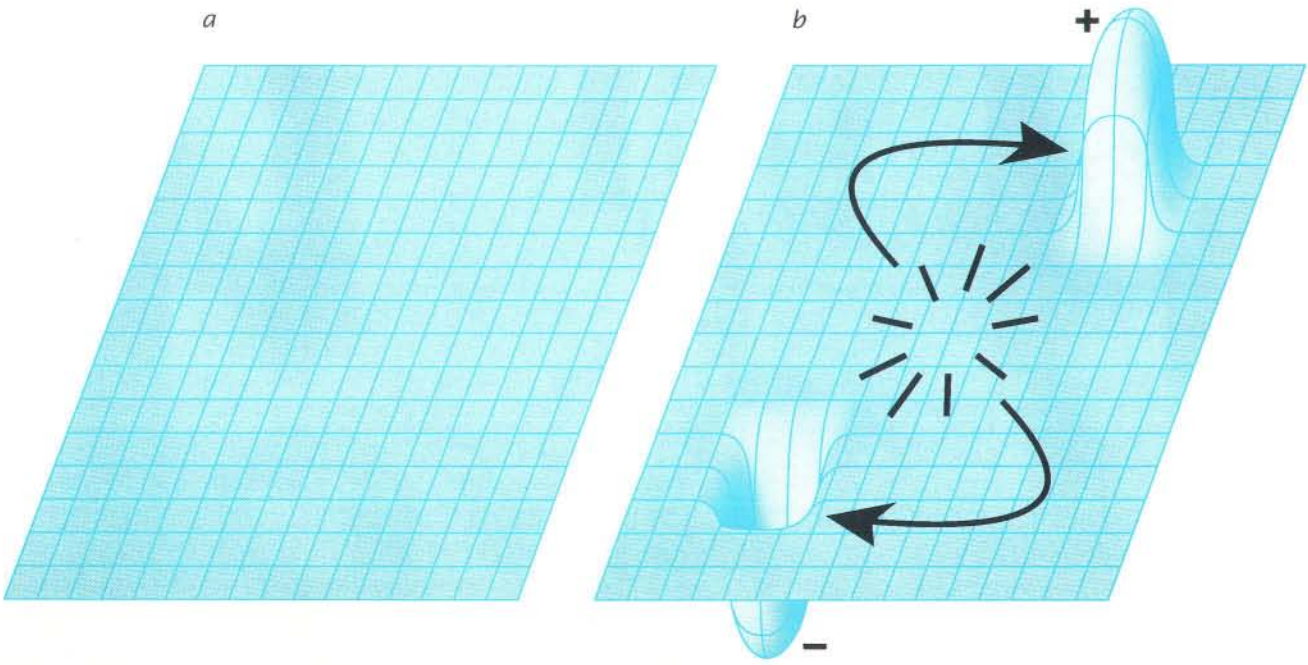
The gravitational field of the earth, for instance, is produced by its rest energy, which equals the mass of the earth multiplied by the square of the speed of light (as given by the famous formula $E = mc^2$). The gravitational field produces a small distortion in the space-time geometry near the earth, resulting in the attractive force that pulls us all toward the ground. In general relativity the energy density of the vacuum has an absolute meaning, and it can be determined by measuring the gravitational field produced not by matter but by the vacuum itself.

Of course, determining the energy density of the vacuum is tantamount to determining the cosmological constant, since one is proportional to the other. It turns out that the cosmological constant can be assigned units of 1 over distance squared. In other words, the square root of the reciprocal of the cosmological constant is a distance. This distance has a direct physical meaning. It is the length scale over which the gravitational effects of a nonzero vacuum energy density would have an obvious and highly visible effect on the geometry of space and time. By studying the geometric properties of the universe over length scales on the order of that distance, the value of the cosmological constant can be measured.

Physicists have been struggling with the issue of the cosmological constant for more than 70 years. The constant was first introduced by Einstein in 1917 in an attempt to eliminate two "problems" in his original formulation of the general theory of relativity. First he thought that without a cosmological constant the general theory could not account for a homogeneous and isotropic universe: one that looks much the same everywhere. (It is remarkable that Einstein even cared about such matters in 1917, since at the time there was no evidence that the universe was homogeneous and isotropic, which indeed it is.) Unfortunately, Einstein's reasoning was incorrect. In 1922 Alexander A. Friedmann showed that the general theory does allow for a homogeneous and isotropic universe, although not a static one: the universe must be expanding (or contracting). Subsequent astronomical observations have convincingly demonstrated that models based on Friedmann's work accurately describe the large-scale structure of the universe.

Einstein was also dissatisfied with his original formulation because the theory did not provide an explanation of inertia. He believed that by adding a cosmological constant he might produce a theory capable of relating the inertial properties of matter directly to the distribution of energy and momentum in the universe, in a manner first suggested by the Austrian physicist and philosopher Ernst Mach. The hope was dashed soon after Einstein's paper appeared by an argument advanced by the Dutch physicist Willem de Sitter, who discovered the space-time we shall discuss.

After such an ignominious start it is not surprising that in 1923 Einstein wrote, perhaps somewhat bitterly, "away with the cosmological term." As we shall see, it has not been so easy to eliminate the cosmological constant—it has survived to frustrate many theoretical physicists since Einstein. George Gamow has written that Einstein felt "the introduction of the cosmological term was the biggest blunder he ever made in his life," but once it had been introduced by Einstein, "the cosmological constant…rears its ugly head again and again."



QUANTUM FLUCTUATIONS are among the phenomena that contribute to the energy density of the vacuum (*a*). According to quantum mechanics, the values of physical quantities tend to fluctuate unavoidably. Hence, pairs of so-called virtu-

At the present time we would appear to be in an excellent position to address the issue of the cosmological constant, because we possess one of the most successful physical theories ever developed, namely, the Standard Model. The Standard Model is the rather unimaginative name given to a collection of theories that successfully describes all the known elementary particles and their interactions. The remarkable ability of the model to interpret and predict the results of an enormous range of particle physics experiments leaves it unchallenged as a model for particle physics (at least up to the highest energies accessible to current particle accelerators).

The Standard Model is a quantum field theory. This means that for every distinct type of fundamental particle in nature there exists a corresponding field in the model used to describe the properties and interactions of that particle. Thus, in the Standard Model there is an electron field, a field for the photon (the electromagnetic field) and a field for each of the known particles.

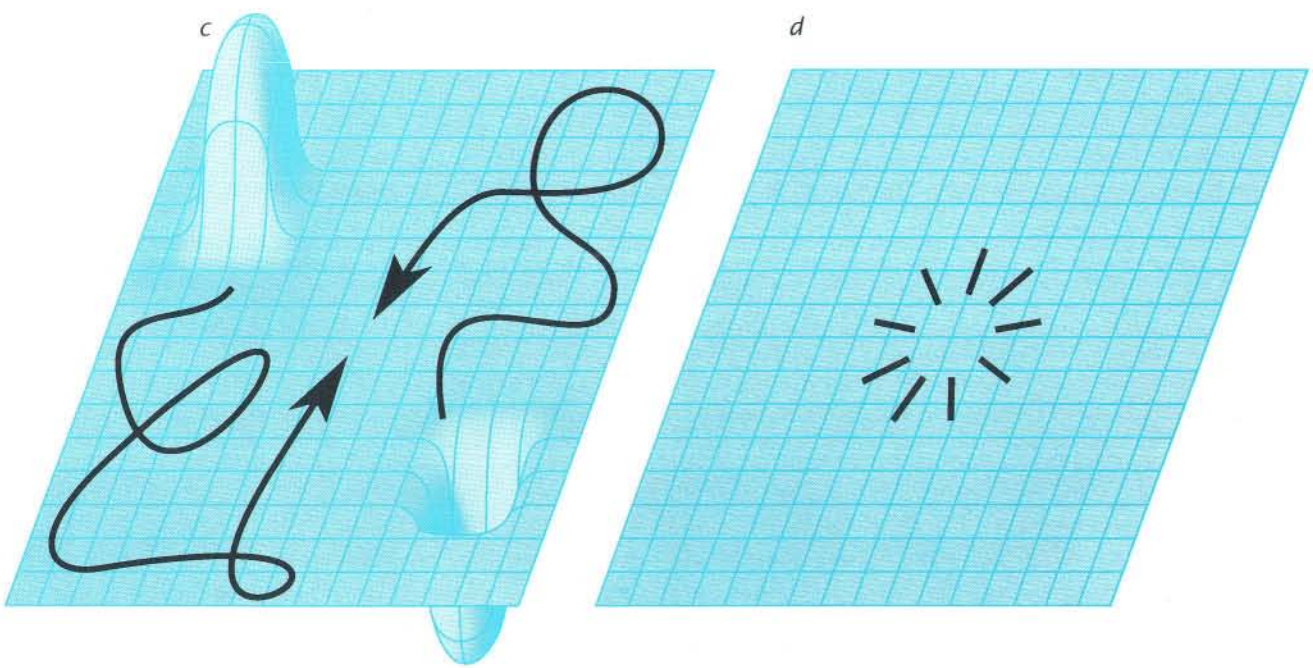
The model depends on a fairly large number of free parameters: numbers that must be determined by experiment and fed into the theory before definite predictions can be made. Examples of free parameters include the values of the masses of the particles and numbers characterizing the strengths of their interactions. Once the numbers have been determined, the model can be used to predict the results of further experiments, and it can be tested on the basis of its predictions. In the past, such tests have been spectacularly successful.

The free parameters of the Standard Model will play a central role in our discussion. Although the model is highly successful, the fact that it depends on such a large number of free parameters seriously limits its predictive power. The model, for example, predicts that an additional particle called the top quark remains to be discovered but is unable to provide a value for its mass, because this is another free parameter of the theory. A key challenge in particle physics today is to develop a more powerful theory based on a smaller number of free parameters that nonetheless incorporate all the successes of the Standard Model. Such a theory would be able to determine the values of some of the parameters that cannot be predicted by the model. In their search for such a theory, physicists are constantly looking for relations among the parameters of the Standard Model that might reveal a deeper structure. As we shall see, the cosmological constant will provide us with such a relation, but in this case we shall get more than we bargained for.

In the Standard Model, as in any quantum field theory, the vacuum is defined as the state of lowest energy or, more properly, as the state of least energy density. This does not imply that the energy density of the vacuum is zero, however. The energy density can in fact be positive, negative or zero depending on the values of various parameters in the theory. Regardless of its value, there are many complex processes that contribute to the total vacuum energy density.

In essence, the total energy density of the vacuum is the sum of three types of terms. The first of the terms is the bare cosmological constant: the value the cosmological constant would have if none of the known particles existed and if the only force in the universe were gravity. The bare cosmological constant is a free parameter that can be determined only by measuring experimentally the true value of the cosmological constant.

The second type of contribution to the total energy density of the vacuum arises in part from quantum fluctuations. The fields in the Standard Model, such as the electron field, experience fluctuations even in the vacuum. Such fluctuations manifest themselves as pairs of so-called virtual particles, which appear spontaneously, briefly interact and then disappear. (Each pair of virtual particles consists of a particle and its corresponding antiparticle, such as the electron and the positron, which have identical masses but oppo-



al particles can appear spontaneously (*b*), interact briefly (*c*) and then disappear (*d*). Here fluctuations are depicted in an abstract and symbolic manner. Each pair of virtual particles consists of a particle and a corresponding antiparticle.

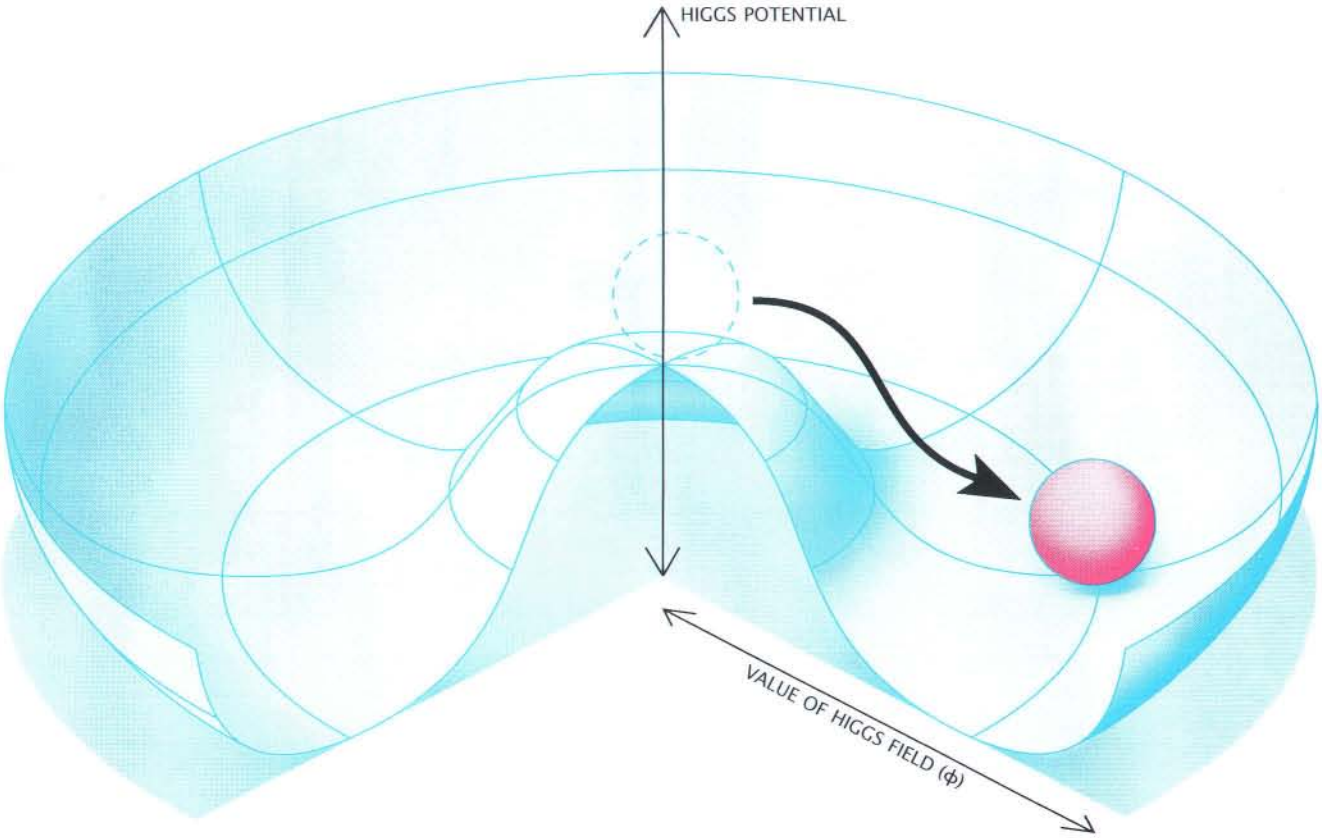
site electric charges.) Although virtual particles cannot be detected by a casual glance at empty space, they have measurable impacts on physics, and in particular they contribute to the vacuum energy density. The contribution made by vacuum fluctuations in the Standard Model depends in a complicated way on the masses and interaction strengths of all the known particles.

The second type of term also depends on at least one additional field—the Higgs field, which represents a massive particle, the Higgs boson, that has not yet been detected. The Higgs field should have a particularly dramatic effect on the energy density of the vacuum state [see "The Higgs Boson," by Martinus J. G. Veltman; SCIENTIFIC AMERICAN, November 1986].

The last type of term that must be included is essentially a fudge factor representing the contributions to the vacuum energy density from additional particles and interactions that may exist but up to now have not been discovered. The value of this term is of course unknown.

The cosmological constant is determined by adding together the three terms we have discussed. Our ability to predict its value using the Standard Model is frustrated by the existence of the bare cosmological constant—a free parameter that can be determined only by carrying out the very measurement we are attempting to predict—and by the sensitivity of the vacuum energy to unknown physics. All is not lost, however, at least not yet. Although all the terms that go into making up the cosmological constant depend in a complicated way on all the parameters of the Standard Model, the values of many of the terms can be fairly accurately estimated. The constituents of protons and neutrons, the "up" and "down" quarks, contribute an amount of about $1/(1 \text{ kilometer})^2$ to the cosmological constant, for instance, and the Higgs field contributes an even larger amount, roughly $1/(10 \text{ centimeters})^2$.

Each of the terms that contributes to the cosmological constant depends on the parameters of the Standard Model in a distinct and independent way. If we assume that the parameters of the model are really free and independent (an assumption we are continually checking in our search for deeper structure), it seems unlikely that these apparently unrelated terms would cancel one another. As a consequence, it seems reasonable to assume that the total cosmological constant will be at least as large or larger than the individual terms we can compute. Such an argument is too crude to predict whether the cosmological constant should be positive or negative, but we would conservatively estimate that its magnitude should be at least $1/(1 \text{ kilometer})^2$, that it could well be something on the order of $1/(10 \text{ centimeters})^2$ and perhaps that it is even larger. In other words, we expect the gravitational effects of a nonzero vacuum energy density to appear as distortions in

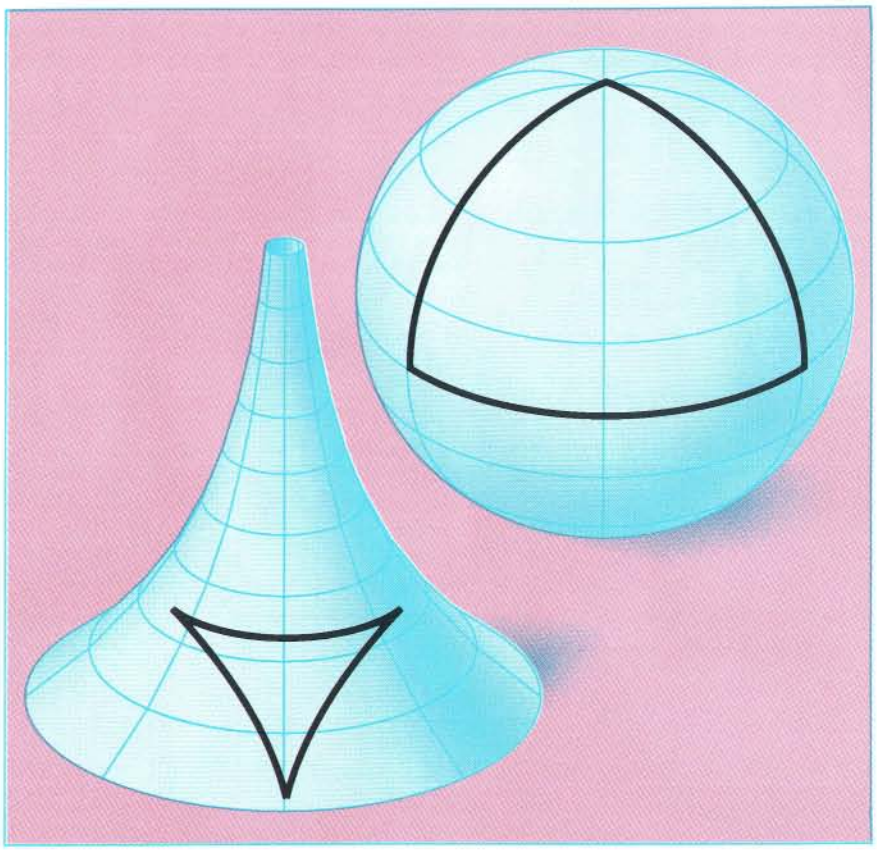


HIGGS FIELD, if it exists, would make a particularly large contribution to the energy density of the vacuum. The Higgs field is the conjectured field corresponding to the particle called the Higgs boson, which is thought to give rise to particle masses. Here the Higgs potential—the part of the vacuum energy density that depends on the value of the Higgs field—is plotted against the value of the field, ϕ. Although the Higgs potential is symmetric about the vertical axis, the vacuum must break the symmetry by choosing a certain position in the trough (*ball*). This is called spontaneous symmetry breaking; it plays a key role in the Standard Model, the theory that describes elementary particles and their interactions.

space-time geometry over distances of one kilometer or less.

It does not require any sophisticated experimentation to show that the theoretical estimate we have just given is wildly wrong. We all know that ordinary Euclidean geometry provides a perfectly adequate description of space over distances much greater than one kilometer. While walking around the block, none of us has ever noticed large distortions in the space-time structure of our neighborhood. If the magnitude of the cosmological constant were as large as our Standard Model estimate, ordinary Euclidean geometry would not be valid over distance scales of one kilometer or even less. If the cosmological constant were negative with a magnitude of $1/(1\ \text{kilometer})^2$, then the sum of the angles of a triangle with sides on the order of one kilometer would be significantly less than 180 degrees, and the volume of a sphere of radius one kilometer would be significantly greater than $4\pi/3$ cubic kilometers.

A positive cosmological constant of order $1/(1\ \text{kilometer})^2$ would have even more bizarre consequences. If the cosmological constant were that large, we would not be able to see objects more than a few kilometers away from us because of the tremendous distortions in space-time structure. In addition, if we walked farther than a few kilometers away from home to see what the rest of the world looked like, the gravitational distortion of space-time would be so great that we could never return home no matter how hard we tried.



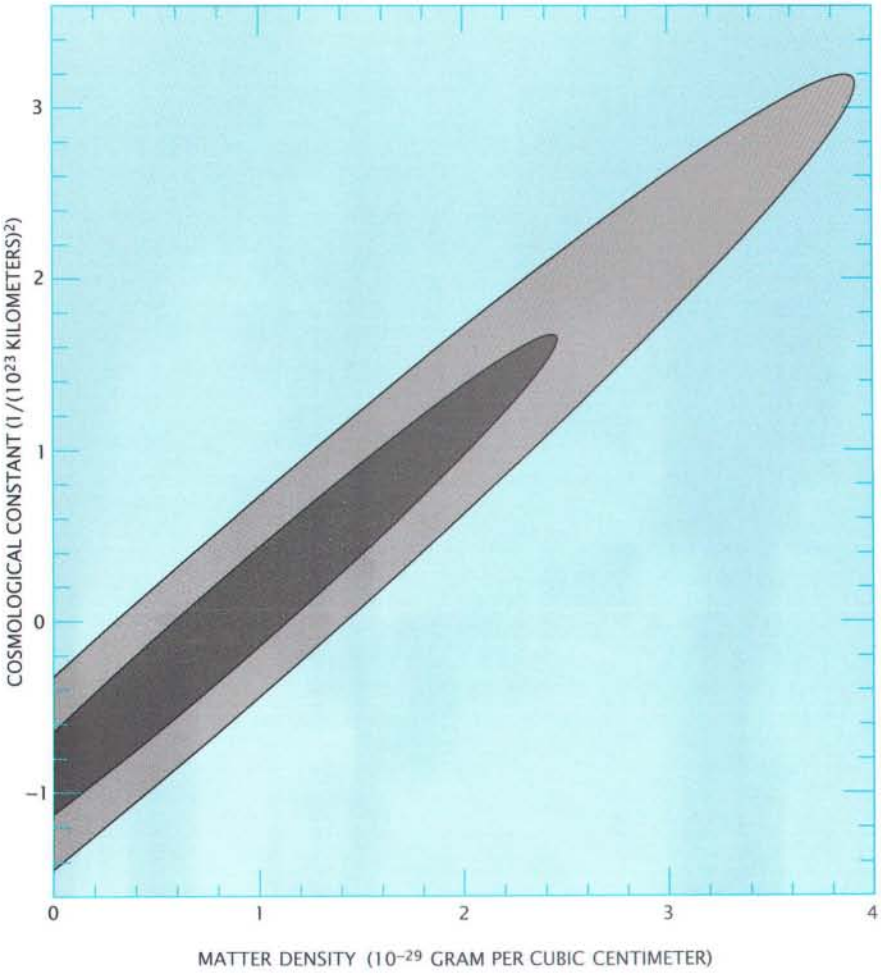
GEOMETRIC DISTORTIONS produced by a nonzero cosmological constant can affect both space and time. Here the effect on spatial geometry is shown, assuming that the distortions are independent of time. A negative cosmological constant would produce a space with negative constant curvature (*left*); a positive constant would produce positive constant curvature (*right*). (The positive case corresponds to the illustration on page 73.) In a space that had a negative curvature, the sum of the angles of a triangle would be less than 180 degrees; with positive curvature, the sum of the angles would be greater than 180.

What if the cosmological constant is nonzero but quite small? In this case we would have to look over large distances to see its effects on space-time structure. Of course, we cannot draw triangles the size of the universe and measure their angles, but we can observe the positions and motions of distant galaxies. By carefully charting the distribution and velocities of distant galaxies, astronomers can deduce the geometric structure of the space-time in which they exist and move.

It has long been recognized that the dominant source of gravitational distortion in the space-time geometry of the universe at large scales appears to be the energy density of matter and not that of the vacuum. Although the energy density of matter and that of the vacuum both affect the geometric structure of the universe, they do so in different and distinguishable ways. Numerous observations have shown that the galaxies in the universe are moving away from one another, a fact that is one of the cornerstones of the expanding universe in the "big bang" cosmology currently accepted. The ordinary gravitational attraction among galaxies tends to slow this expansion. As the galaxies get farther away from one another, their gravitational attraction weakens, and so the rate at which the expansion slows decreases with time. Thus, the effect of ordinary matter on the expansion of the universe is to decelerate the expansion at an ever decreasing rate.

What effects would a nonzero cosmological constant have on the expansion rate of the universe? A negative cosmological constant would tend to slow the expansion of the galaxies, but at a rate that is constant, not decreasing with time. A positive cosmological constant, on the other hand, would tend to make the galaxies accelerate away from one another and increase the expansion rate of the universe. Comprehensive studies of the expansion rates of distant galaxies show no evidence for either a positive or a negative cosmological constant.

A good example of how astronomers can measure the geometry of the universe and look for a nonzero cosmological constant is provided by the recently published work of Edwin D. Loh and Earl J. Spillar of Princeton University. Their survey counts the numbers of galaxies in regions of a specific size at various locations in space. If we assume that on the average the number of galaxies per unit volume is the same everywhere, then by counting galaxies in a region we are estimating the volume of that region. By measuring volumes of regions far from us, we are determining the relation between distance and volume over very large



COSMOLOGICAL CONSTANT has been probed by counting the number of galaxies in regions of the universe and thereby determining the geometry of those regions. The graph plots allowed values of the cosmological constant versus the matter density of the universe. (The black area corresponds to values that are allowed with a confidence of 67 percent; the gray area is a region of 95 percent confidence.) The units are approximate, but the graph shows that the magnitude of the cosmological constant must be less than about $1/(10^{23}$ kilometers$)^2$, some 46 orders of magnitude smaller than the value predicted on the basis of the Standard Model. The graph is from an analysis by Edwin D. Loh of Princeton University, based on work with Earl J. Spillar, also of Princeton.

scales and at earlier times, since the light from distant galaxies takes a long time to reach us—billions of years in the case of this survey.

Although such surveys contain many subtle sources of potential error, the results differ so startlingly from our theoretical estimate that errors of a factor of two or even 10 are fairly insignificant. All galactic surveys agree that there is no evidence for any spacetime distortions resulting from a nonvanishing cosmological constant out to the farthest distances accessible to astronomers, about 10 billion light-years, or 10^{23} kilometers. This implies that the magnitude of the cosmological constant must be smaller than $1/(10^{23}$ kilometers$)^2$. Our theoretical estimate suggesting a magnitude larger than $1/(1$ kilometer$)^2$ is incorrect by, at the least, an astonishing factor of 10^{46}. Few theoretical estimates in the history of physics made on the basis of what seemed to be reasonable assumptions have ever been so inaccurate.

The stupendous failure we have experienced in trying to predict the value of the cosmological constant is far more than a mere embarrassment. Recall that the basic assumption we used to obtain our estimate was that there are no unexpected cancellations among the various terms in the sum determining the total energy density of the vacuum. This expectation was based on the assumed independence of the free parameters of the Standard Model. Clearly, this assumption is spectacularly wrong. There must in fact be a miraculous conspiracy occurring among both the known and the unknown parameters governing particle physics, with the result that the many terms making up the cosmological constant add up to a quantity more than 46 orders of magnitude smaller than the individual terms in the sum. In other words, the small value of the cosmological constant is telling us that a remarkably precise and totally unexpected relation exists among all the parameters of the Standard Model, the bare cosmological constant and unknown physics.

A relation among the free parameters of the Standard Model is just what we seek in our quest to discover deeper and more predictive theories. How could such a complex relation among what we thought were free and unconstrained parameters arise, and what does it mean?

In answering this question, it is well to keep in mind two examples from an earlier period in the history of physics. In the mid-19th century the speed of light had been measured, and theories existed describing electric and magnetic phenomena, but it had not yet been shown that light propagation is an electromagnetic effect. Several physicists noticed, however, a curious relation between the speed of light and two parameters that enter into the equations for electric and magnetic phenomena. In modern notation what they noticed was that the electromagnetic permittivity constant ε_0 and the magnetic permeability constant μ_0 could be incorporated into the formula $\sqrt{(1/\varepsilon_0\mu_0)}$, yielding a quantity that is numerically equal to the measured velocity of light (at least within the rather large experimental errors of that time).

The workers appreciated the fact that this was either a miraculous numerical coincidence or evidence of a fundamental and as yet undiscovered relation between electromagnetic phenomena and light. James Clerk Maxwell was also aware of this numerical curiosity, and it served as an important inspiration for him in showing, through the set of equations now bearing his name, that the propagation of light is indeed profoundly related to electric and magnetic phenomena.

Does the remarkable relation among the parameters of the Standard Model implied by the small value of the cosmological constant suggest that a wonderful unifying theory awaits our discovery? Before jumping to such a conclusion, I should like to relate another example from the history of electromagnetic theory.

After Maxwell had incorporated light propagation into electromagnetic theory, it was generally assumed that light waves traveled through a medium known as the ether. Using an interferometer, Albert A. Michelson and Edward W. Morley attempted to measure the velocity of the earth as it traveled through the ether. They found that the relative velocity was zero: the velocity of the earth and the velocity of the ether were identical. This is another relation involving what was then thought to be a fundamental parameter of nature, namely, the velocity of the ether. Did the discovery point the way to a unified theory relating a fundamental property of electromagnetism to the motion of the earth?

Although the idea that the ether drifted with the earth was suggested, the zero result of the Michelson-Morley experiment is actually explained by Einstein's special theory of relativity, which showed that the conception of the ether being used in that era was inconsistent with the symmetries of space and time. No theory providing a fundamental relation between the velocity of the ether and something as idiosyncratic as the velocity of the earth has survived. That is hardly surprising. The velocity of the earth is affected by many things—the shape and size of its orbit around the sun, the mass of the sun and the motion of the sun in the galaxy, for instance—that seem completely unrelated to issues in the theory of electromagnetism. There is no fundamental relation between the velocity of the ether and the velocity of the earth because the ether itself as the 19th-century theorists imagined it does not even exist.

In both examples a surprising relation between parameters of nature foreshadowed dramatic and revolutionary new discoveries. We have every reason to believe the mysterious relation implied by the vanishingly small value of the cosmological constant indicates that discoveries as important as these remain to be made. The two examples we have considered are quite different. The first relation, which involves two parameters of electromagnetism and one from light propagation, is what physicists today would call a "natural" relation: one that involves a small number of well-known parameters. The existence of a natural relation may indicate that a unifying theory exists, and, more important, it suggests that such a theory can be discovered.

The second example, in which the velocity of the ether was related to the velocity of the earth, is what today would be called an "unnatural" relation: one that involves many parameters, some of which are unknown or even unknowable. It seems unlikely, for instance, that we will ever know and understand all the many factors that determine what the velocity of the earth is in relation to the distant galaxies. Any unified theory developed to account for an unnatural relation would have to explain the values of many known and unknown parameters all at once. It seems quite unlikely that such a theory could be discovered even if it did exist.

Our example indicates that an unnatural relation suggests a deep misunderstanding about the essence of what is being measured and related, rather than the existence of an underlying unified theory. As a consequence, an unnatural relation may point to an even more dramatic revolution in our thinking than a natural one would.

If we discount the possibility that the vanishingly small value of the cosmological constant is accidental, we must accept that it has profound implications for physics. Before we launch into constructing new unified models, however, we must face the dilemma that the relation implied by the vanishing of the cosmological constant is unnatural. The miraculous cancellations required to produce an acceptably small cosmological constant depend on all the parameters relevant to particle physics, known and unknown. To predict a zero (or small) value for the cosmological constant, a unified theory would face the imposing task of accounting for every parameter affecting particle physics. Even worse, achieving a sufficiently small cosmological constant requires that extremely precise (one part in 10^{46} or more) cancellations take place; the parameters would have to be predicted by the theory with extraordinary accuracy before any improvement in the situation regarding the cosmological constant would even be noticeable. Constructing such a theory, even if it does exist, seems to be an awesome if not impossible task.

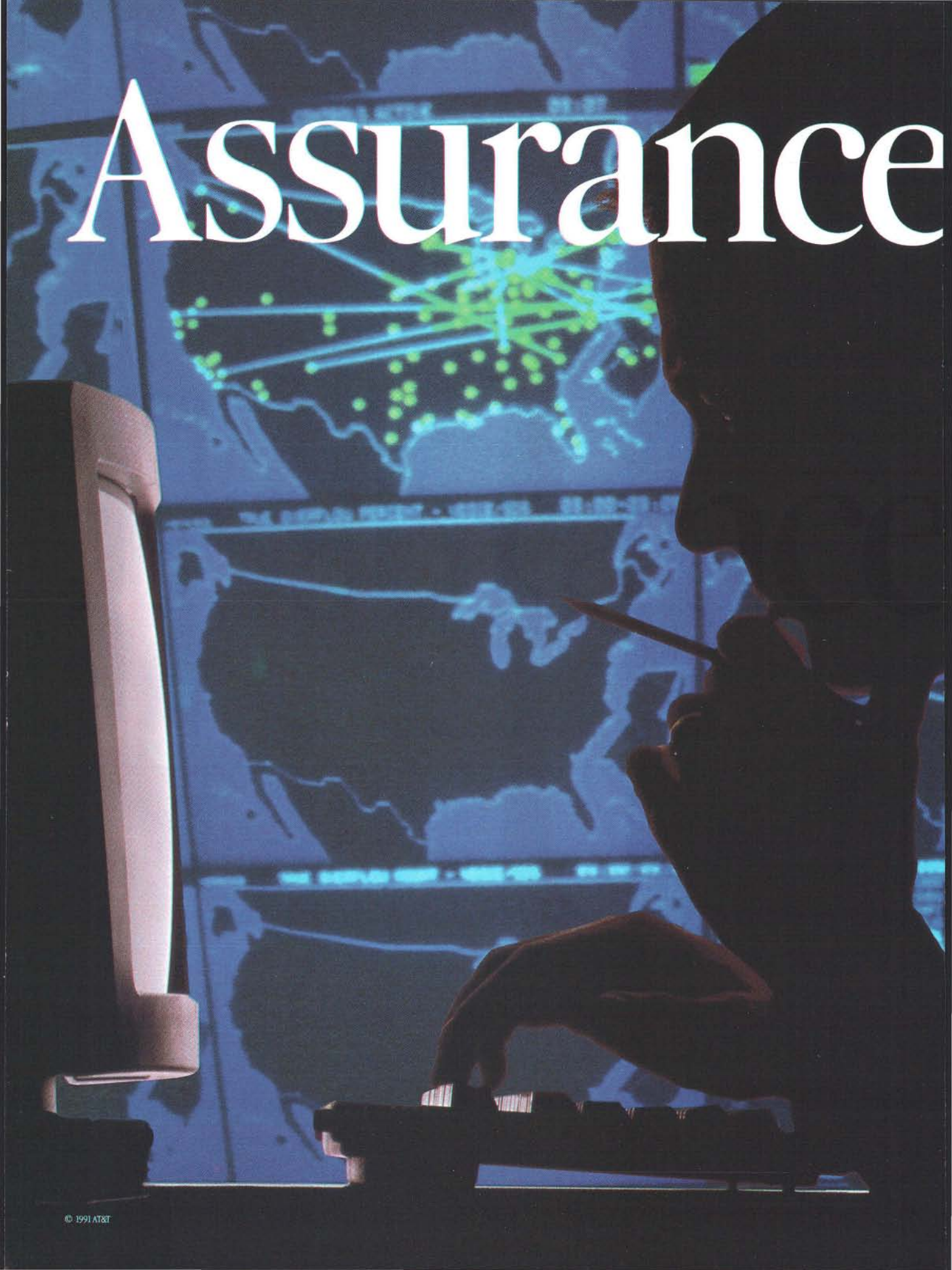
Although certain theories of the "ether drift" variety have been proposed, most efforts concerning the cosmological constant now focus on finding the underlying misunderstanding, the missing piece of the Standard Model or the misconception about the vacuum, which once understood will either eliminate the problem or at least turn it into a natural one. As long as the problem of the cosmological constant remains unnatural, the only hope we have for finding a solution is to stumble on an all-encompassing theory capable of accounting for all particle physics parameters with nearly perfect accuracy. If we can change the relation required to produce an acceptably small vacuum energy density into a natural one, then, even though we have not yet accounted for its value, we at least reduce the issue of the cosmological constant to a more manageable problem involving a reasonable number of known parameters that only have to be predicted with a moderate degree of accuracy. There is little to report to date about this effort. In spite of a lot of hard work and creative ideas, we still do not know why the cosmological constant is so small.

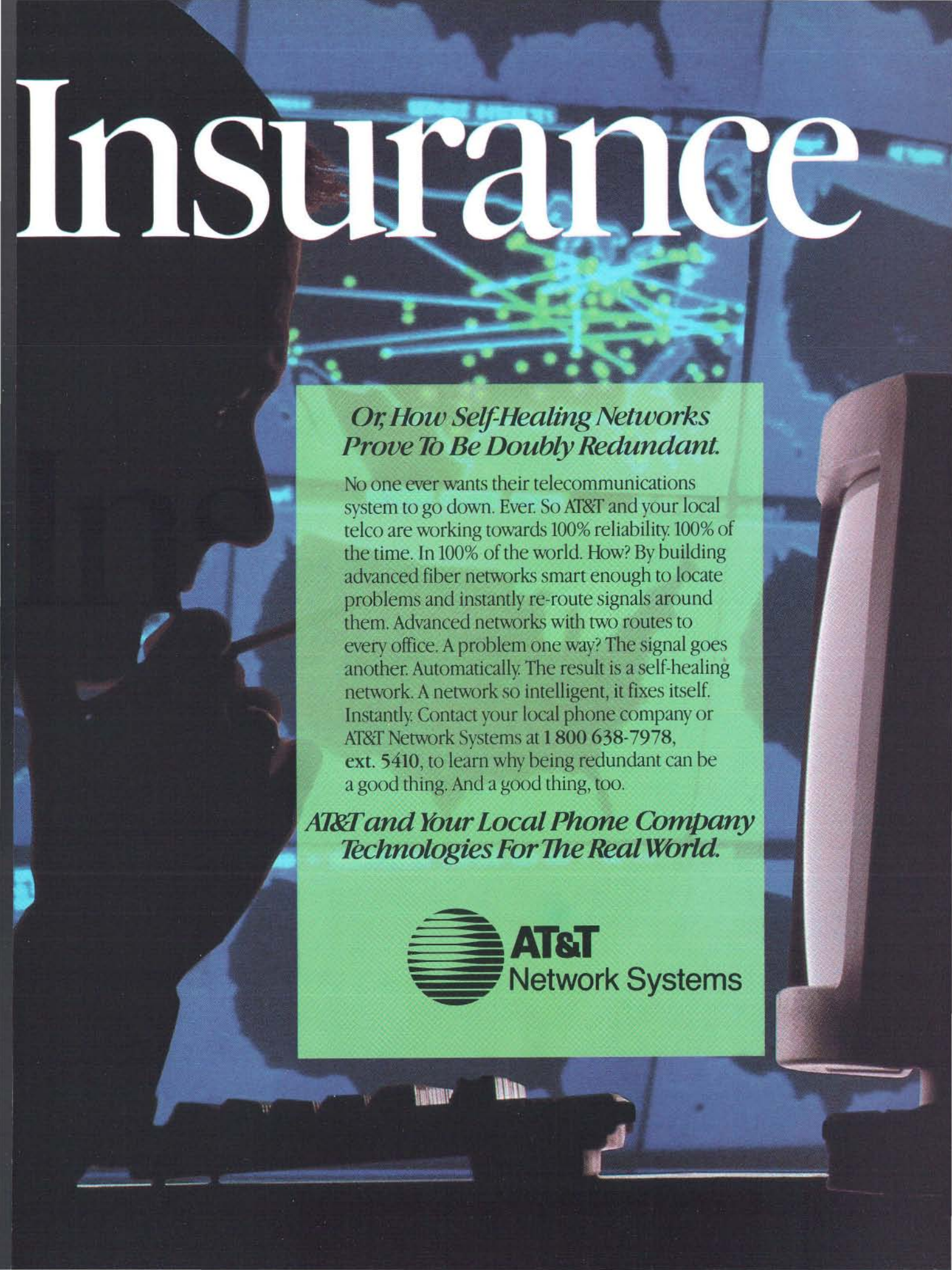
Even though nature does not, in the words of Aristotle, "abhor a vacuum," perhaps it does abhor a vacuum that is not empty. By introducing the ether in the early days of electromagnetic theory, Maxwell and others cluttered the vacuum with a hypothetical fluid that had complex properties. Michelson and Morley showed that this view of the vacuum was inconsistent with experimental reality, and Einstein showed that it was inconsistent with the symmetries of the universe.

Quantum field theories also fill the emptiness of the vacuum, this time with quantum fluctuations and fields rather than ether. These modern forms of clutter are consistent with the special theory of relativity, but they seem to cause problems when they are viewed in the framework of the general theory. With the mystery of the cosmological constant, perhaps we are again paying the price for dumping too much into the vacuum. The Standard Model, which has a large number of fluctuating quantum fields, including a Higgs field, is a particularly egregious polluter of the vacuum. There is no doubt that the resulting theory is a beautiful and highly successful structure, but it may be based on a conception of the vacuum or of space-time that is flawed. It is our challenge to repair that faulty foundation without destroying the towering edifice we have built on it.

FURTHER READING

EXPANDING UNIVERSES. Erwin Schrödinger. Cambridge University Press, 1956.
THE LARGE-SCALE STRUCTURE OF SPACE-TIME. S. W. Hawking and G.F.R. Ellis. Cambridge University Press, 1973.
THE ACCIDENTAL UNIVERSE. P.C.W. Davies. Cambridge University Press, 1982.
'SUBTLE IS THE LORD...': THE SCIENCE AND THE LIFE OF ALBERT EINSTEIN. Abraham Pais. Oxford University Press, 1982.




Insurance
Or, How Self-Healing Networks Prove To Be Doubly Redundant.
No one ever wants their telecommunications system to go down. Ever. So AT&T and your local telco are working towards 100% reliability. 100% of the time. In 100% of the world. How? By building advanced fiber networks smart enough to locate problems and instantly re-route signals around them. Advanced networks with two routes to every office. A problem one way? The signal goes another. Automatically. The result is a self-healing network. A network so intelligent, it fixes itself. Instantly. Contact your local phone company or AT&T Network Systems at 1 800 638-7978, ext. 5410, to learn why being redundant can be a good thing. And a good thing, too.
AT&T and Your Local Phone Company Technologies For The Real World.
AT&T
Network Systems

The Structure of the Hereditary Material

An account of the investigations which have led to the formulation of an understandable structure for DNA. The chemical reactions of this material within the nucleus govern the process of reproduction

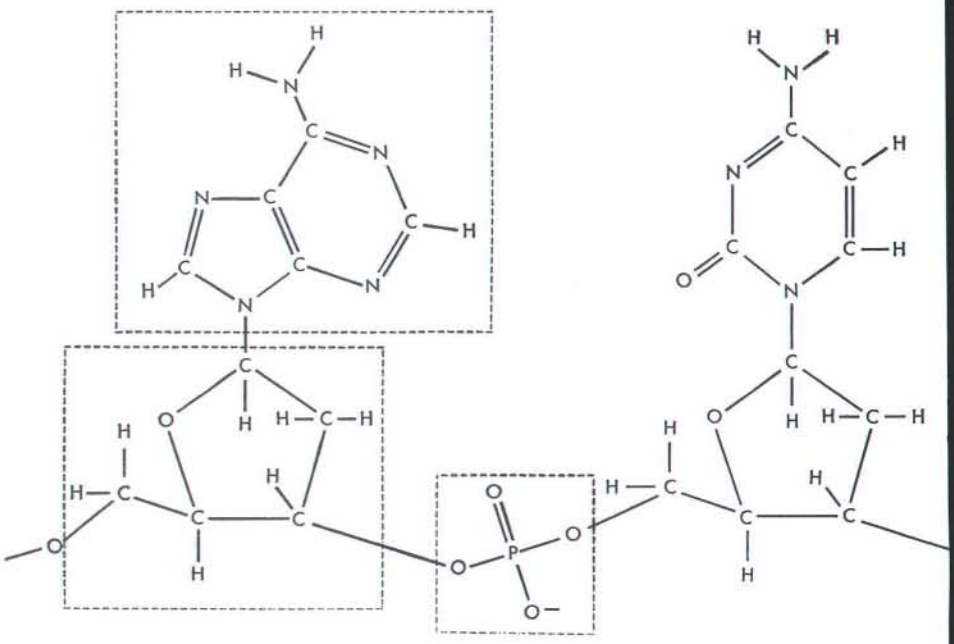
by F.H.C. Crick

Viewed under a microscope, the process of mitosis, by which one cell divides and becomes two, is one of the most fascinating spectacles in the whole of biology. No one who watches the event unfold in speeded-up motion pictures can fail to be excited and awed. As a demonstration of the powers of dynamic organization possessed by living matter, the act of division is impressive enough, but even more stirring is the appearance of two identical sets of chromosomes where only one existed before. Here lies biology's greatest challenge: How are these fundamental bodies duplicated? Unhappily, the copying process that takes place in mitosis is beyond the resolving power of microscopes, but much is being learned about it in other ways.

One approach is the study of the nature and behavior of whole living cells; another is the investigation of substances extracted from them. This article will discuss only the second approach, but both are indispensable if we are ever to solve the problem; indeed, some of the most exciting results are being obtained by what might loosely be described as a combination of the two methods.

Chromosomes consist mainly of three kinds of chemical: protein, desoxyribonucleic acid (DNA) and ribonucleic acid (RNA). (Since RNA is only a minor component, we shall not consider it in detail here.) The nucleic acids and the proteins have several features in common. They are all giant molecules, and each type has the general structure of a main backbone with side groups attached. The proteins have about 20 different kinds of side groups; the nucleic acids usually only four (and of a different type). The smallness of these numbers itself is striking, for there is no obvious chemical reason

F.H.C. CRICK is a British biologist who was originally trained as a physicist. After war years spent designing mines for the British Admiralty, he decided to go into molecular biology. He obtained a Medical Research Council studentship, entered the Strangeways Laboratory in Cambridge, worked on the viscosity of the cytoplasm of chick fibroblasts and "read everything I could lay my hands on." He then joined a molecular biology unit sponsored by the Medical Research Council, where he was able to concentrate on molecular structure. In 1962 Crick and his colleagues James D. Watson and Maurice Wilkins were awarded the Nobel Prize for Physiology or Medicine for the work described here. Since 1977 Crick has been Kieckhefer Distinguished Professor at the Salk Institute for Biological Studies in San Diego.



FRAGMENT OF CHAIN of DNA shows the three basic units that make up the molecule. Repeated often in a long chain, they make it 1,000 times as long as it is

Reprinted from the October 1954 issue

why many more types of side groups should not occur. Another interesting feature is that no protein or nucleic acid occurs in more than one optical form; there is never an optical isomer, or mirror-image molecule. This shows that the shape of the molecules must be important.

These generalizations (with minor exceptions) hold over the entire range of living organisms, from viruses and bacteria to plants and animals. The impression is inescapable that we are dealing with a very basic aspect of living matter, and one having far more simplicity than we would have dared to hope. It encourages us to look for simple explanations for the formation of these giant molecules.

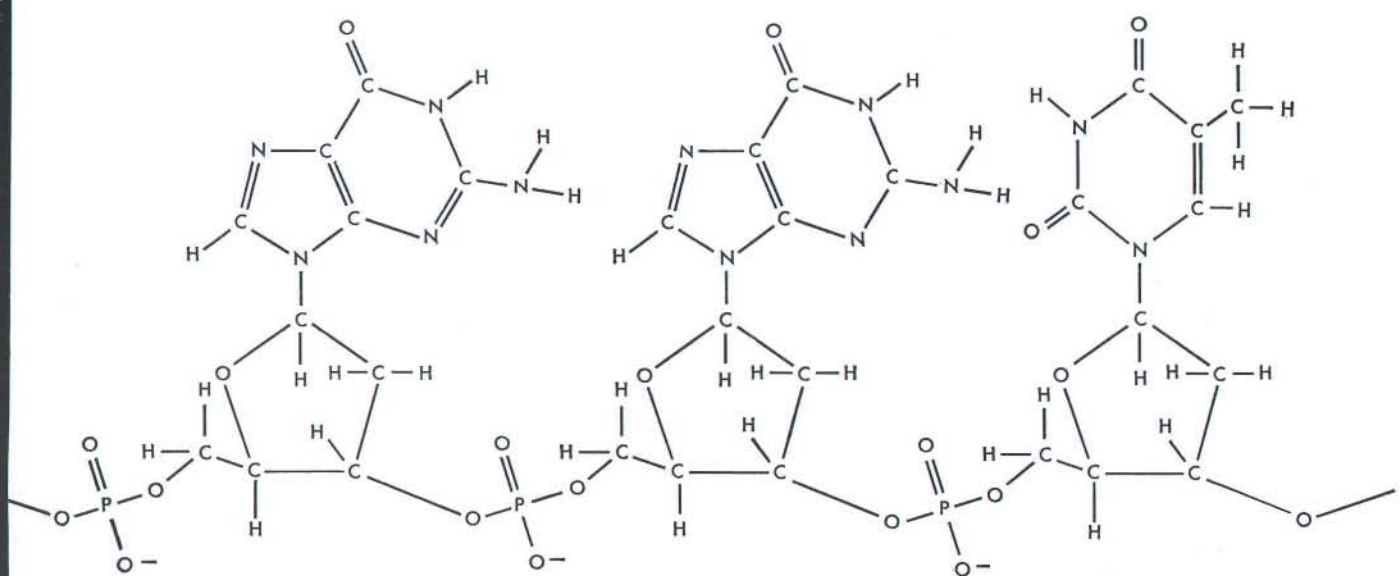
The most important role of proteins is that of the enzymes—the machine tools of the living cell. An enzyme is specific, often highly specific, for the reaction which it catalyzes. Moreover, chemical and X-ray studies suggest that the structure of each enzyme is itself rigidly determined. The side groups of a given enzyme are probably arranged in a fixed order along the polypeptide backbone. If we could discover how a cell produces the appropriate enzymes, in particular how it assembles the side groups of each enzyme in the correct order, we should have gone a long way toward explaining the simpler forms of life in terms of physics and chemistry.

We believe that this order is controlled by the chromosomes. In recent years suspicion has been growing that the key to the specificity of the chromosomes lies not in their protein but in their DNA. DNA is found in all chromosomes—and only in the chromosomes (with minor exceptions). The amount of DNA per chromosome set is in many cases a fixed quantity for a given species. The sperm, having half the chromosomes of the normal cell, has about half the amount of DNA, and tetraploid cells in the liver, having twice the normal chromosome complement, seem to have twice the amount of DNA. This constancy of the amount of DNA is what one might expect if it is truly the material that determines the hereditary pattern.

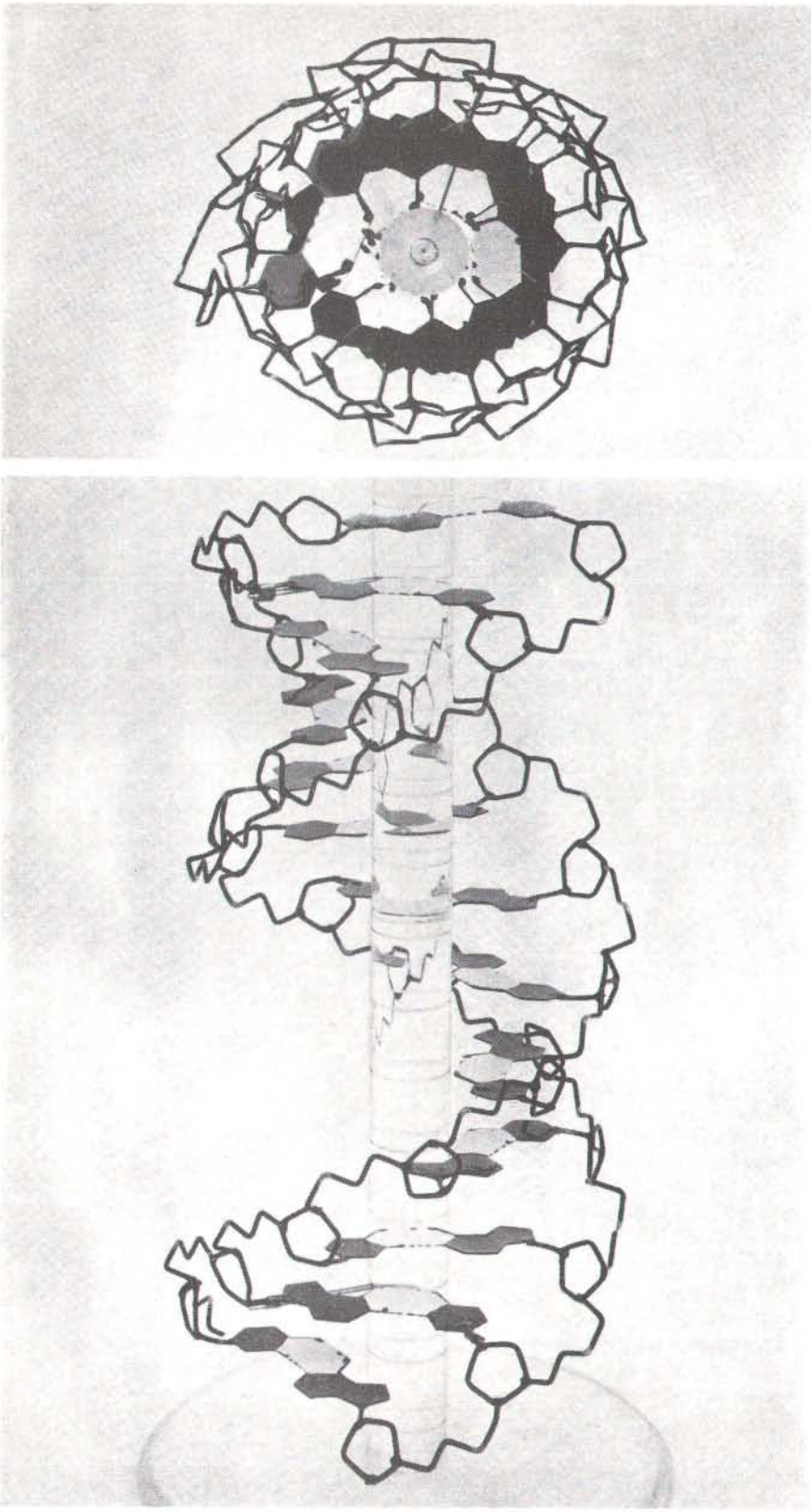
Then there is suggestive evidence in two cases that DNA alone, free of protein, may be able to carry genetic information. The first of these is the discovery that the "transforming principles" of bacteria, which can produce all inherited change when added to the cell, appear to consist only of DNA. The second is the fact that during the infection of a bacterium by a bacteriophage the DNA of the phage penetrates into the bacterial cell while most of the protein, perhaps all of it, is left outside.

DNA can be extracted from cells by mild chemical methods, and much experimental work has been carried out to discover its chemical nature. This work has been conspicuously successful. It is now known that DNA consists of a very long chain made up of alternate sugar and phosphate groups [*see illustration below*]. The sugar is always the same sugar, known as desoxyribose. And it is always joined onto the phosphate in the same way, so that the long chain is perfectly regular, repeating the same phosphate-sugar sequence over and over again.

But while the phosphate-sugar chain is perfectly regular, the molecule as a whole is not, because each sugar has a "base" attached to it and the base is not always the same. Four different types of base are commonly found: two of them are purines, called adenine and guanine, and two are pyrimidines, known as thymine and cytosine. So far as is known the order in which they follow one another along the chain is irregular and probably varies from one piece of DNA to another. In fact, we suspect that the order of the bases



thick. The backbone is made up of pentose sugar molecules (*middle square*), linked by phosphate groups (*bottom square*). The bases (*top square*) adenine, cytosine, guanine and thymine protrude off each sugar in irregular order.



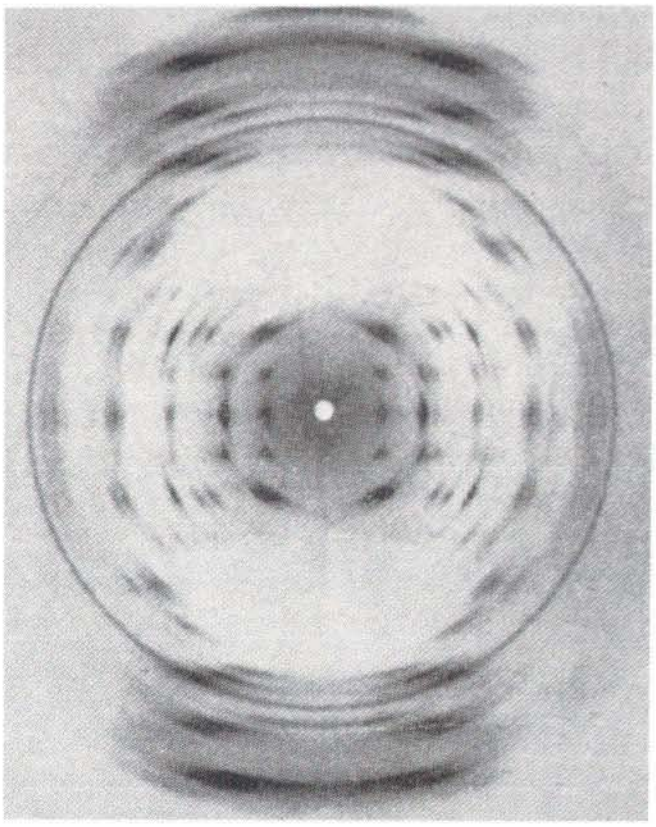
STRUCTURAL MODEL shows a pair of DNA chains wound as a helix about the fiber axis. The pentose sugars can be plainly seen. From every one on each chain protrudes a base, linked to an opposing one at the same level by a hydrogen bond. These base-to-base links act as horizontal supports, holding the chains together. The upper photograph is a view from the top.

is what confers specificity on a given DNA. Because the sequence of the bases is not known, one can only say that the *general* formula for DNA is established. Nevertheless, this formula should be reckoned one of the major achievements of biochemistry, and it is the foundation for all the ideas described in the rest of this article.

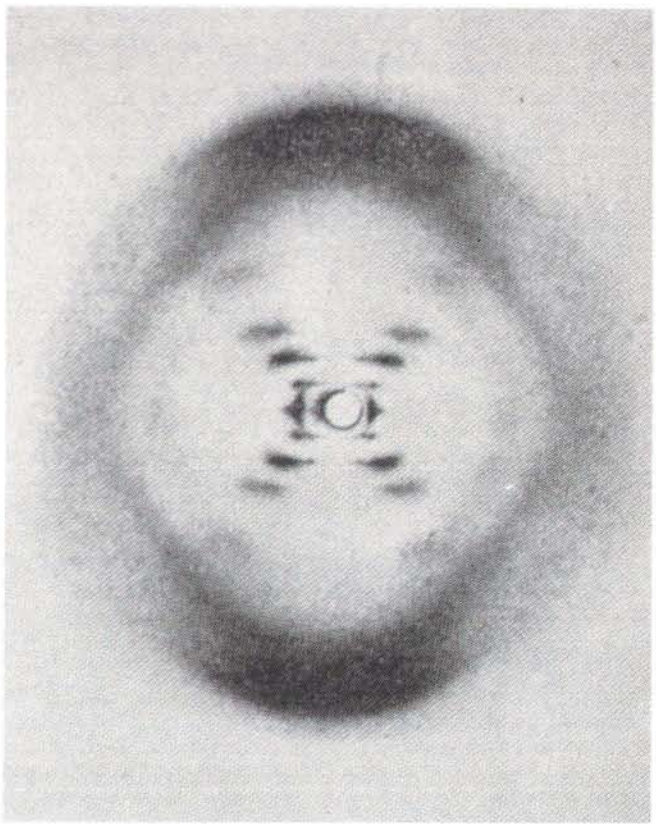
At one time it was thought that the four bases occurred in equal amounts, but in recent years this idea has been shown to be incorrect. E. Chargaff and his colleagues at Columbia University, A. E. Mirsky and his group at the Rockefeller Institute for Medical Research and G. R. Wyatt of Canada have accurately measured the amounts of the bases in many instances and have shown that the relative amounts appear to be fixed for any given species, irrespective of the individual or the organ from which the DNA was taken. The proportions usually differ for DNA from different species, but species related to one another may not differ very much.

Although we know from the chemical formula of DNA that it is a chain, this does not in itself tell us the shape of the molecule, for the chain, having many single bonds around which it may rotate, might coil up in all sorts of shapes. However, we know from physical-chemical measurements and electron microscope pictures that the molecule usually is long, thin and fairly straight, rather like a stiff bit of cord. It is only about 20 angstroms thick (one angstrom equals one 100 millionth of a centimeter). This is very small indeed, in fact not much more than a dozen atoms thick. The length of the DNA seems to depend somewhat on the method of preparation. A good sample may reach a length of 30,000 angstroms, so that the structure is more than 1,000 times as long as it is thick. The length inside the cell may be much greater than this, because there is always the chance that the extraction process may break it up somewhat.

None of these methods tells us anything about the detailed arrangement in space of the atoms inside the molecule. For this, it is necessary to use X-ray diffraction. The average distance between bonded atoms in an organic molecule is about $1\,^1/_2$ angstroms; between unbonded atoms, three to four angstroms. X rays have a small enough wavelength ($1\,^1/_2$ angstroms) to resolve the atoms, but unfortunately, an X-ray diffraction pho-



STRUCTURE *A* is the crystalline form of DNA found at relatively low humidity. This X-ray photograph is by H. R. Wilson.



STRUCTURE *B* is DNA's paracrystalline form. The molecules are less regularly arranged. The picture is by R. E. Franklin.

tograph is not a picture in the ordinary sense of the word. We cannot focus X rays as we can ordinary light; hence, a picture can be obtained only by roundabout methods. Moreover, it can show clearly only the periodic, or regularly repeated, parts of the structure.

With patience and skill, several English workers have obtained good diffraction pictures of DNA extracted from cells and drawn into long fibers. The first studies, even before details emerged, produced two surprises. First, they revealed that the DNA structure could take two forms. In relatively low humidity, when the water content of the fibers was about 40 percent, the DNA molecules gave a crystalline pattern, showing that they were aligned regularly in all three dimensions. When the humidity was raised and the fibers took up more water, they increased in length by about 30 percent, and the pattern tended to become "paracrystalline," which means that the molecules were packed side by side in a less regular manner, as if the long molecules could slide over one another somewhat. The second surprising result was that DNA from different species appeared to give identical X-ray patterns, despite the fact that the amounts of the four bases present varied. This was particularly odd because of the existence of the crystalline form just mentioned. How could the structure appear so regular when the bases varied? It seemed that the broad arrangement of the molecule must be independent of the exact sequence of the bases, and it was therefore thought that the bases play no part in holding the structure together. As we shall see, this turned out to be wrong.

The early X-ray pictures showed a third intriguing fact: namely, the repeats in the crystallographic pattern came at much longer intervals than the chemical repeat units in the molecule. The distance from one phosphate to the next cannot be more than about seven angstroms, yet the crystallographic repeat came at intervals of 28 angstroms in the crystalline form and 34 angstroms in the paracrystalline form; that is, the chemical unit repeated several times before the structure repeated crystallographically.

J. D. Watson and I, working in the Medical Research Council Unit at the Cavendish Laboratory in Cambridge, were convinced that we could get somewhere near the DNA structure by building scale models based on the X-ray patterns obtained by M.H.F. Wilkins, Rosalind Franklin and their coworkers at King's College, London. A great deal is known about the exact distances between bonded atoms in molecules, about the angles between the bonds and about the size of atoms—the so-called van der Waals distance between adjacent nonbonded atoms. This information is easy to embody in scale models. The problem is rather like a three-dimensional jigsaw puzzle with curious pieces joined together by rotatable joints (single bonds between atoms).

To get anywhere at all, we had to make some assumptions. The most important one had to do with the fact that the crystallographic repeat did not coincide with the repetition of chemical units in the chain but came at much longer intervals. A possible explanation was that all the links in the chain were the same, but the X rays were seeing every tenth link, say, from

the same angle and the others from different angles. What sort of chain might produce this pattern? The answer was easy: the chain might be coiled in a helix. (A helix is often loosely called a spiral; the distinction is that a helix winds not around a cone but around a cylinder, as a winding staircase usually does.) The distance between crystallographic repeats would then correspond to the distance in the chain between one turn of the helix and the next.

We had some difficulty at first because we ignored the bases and tried to work only with the phosphate-sugar backbone. Eventually we realized that we had to take the bases into account, and this led us quickly to a structure which we now believe to be correct in its broad outlines.

This particular model contains a pair of DNA chains wound around a common axis. The two chains are linked together by their bases. A base on one chain is joined by very weak bonds to a base at the same level on the other chain, and all the bases are paired off in this way right along the structure. In the diagram on page 84, the two ribbons represent the phosphate-sugar chains, and the pairs of bases holding them together are symbolized as horizontal rods. Paradoxically, in order to make the structure as symmetric as possible, we had to have the two chains run in opposite directions; that is, the sequence of the atoms goes one way in one chain and the opposite way in the other. Thus, the figure looks exactly the same whichever end is turned up.

Now we found that we could not arrange the bases any way we pleased; the four bases would fit into the structure only in certain pairs. In any pair there must always be one big one (purine) and one little one (pyrimidine). A pair of pyrimidines is too short to bridge the gap between the two chains, and a pair of purines is too big to fit into the space.

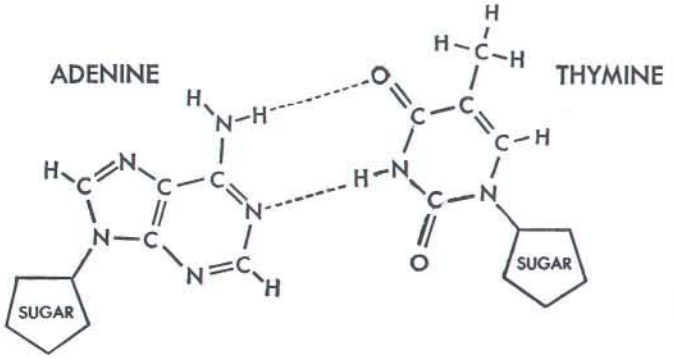
At this point we made an additional assumption. The bases can theoretically exist in a number of forms depending on where the hydrogen atoms are attached. We assumed that for each base one form was much more probable than all the others. The hydrogen atoms can be thought of as little knobs attached to the bases, and the way the bases fit together depends crucially on where these knobs are. With this assumption the only possible pairs that will fit in are adenine with thymine and guanine with cytosine.

The way these pairs are formed is shown in the diagrams below. The dotted lines show the hydrogen bonds, which hold the two bases of a pair together. They are very weak bonds; their energy is not many times greater than the energy of thermal vibration at room temperature. (Hydrogen bonds are the main forces holding different water molecules together, and it is because of them that water is a liquid at room temperatures and not a gas.)

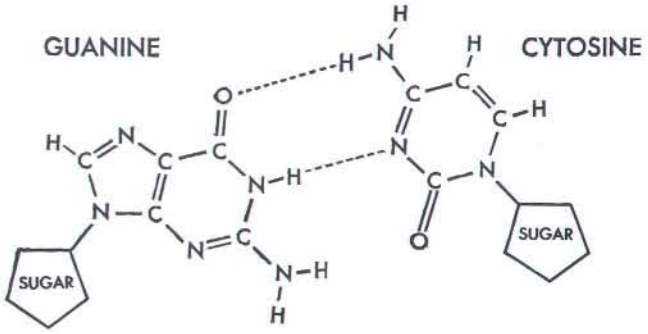
Adenine must always be paired with thymine, and guanine with cytosine; it is impossible to fit the bases together in any other combination in our model. (This pairing is likely to be so fundamental for biology that I could not help wondering whether some day an enthusiastic scientist will christen his newborn twins Adenine and Thymine!) The model places no restriction, however, on the sequence of pairs along the structure. Any specified pair can follow any other. This is because a pair of bases is flat, and since in this model they are stacked roughly like a pile of coins, it does not matter which pair goes above which.

It is important to realize that the specific pairing of the bases is the direct result of the assumption that both phosphate-sugar chains are helical. This regularity implies that the distance from a sugar group on one chain to that on the other at the same level is always the same, no matter where one is along the chain. It follows that the bases linked to the sugars always have the same amount of space in which to fit. It is the regularity of the phosphate-sugar chains, therefore, that is at the root of the specific pairing.

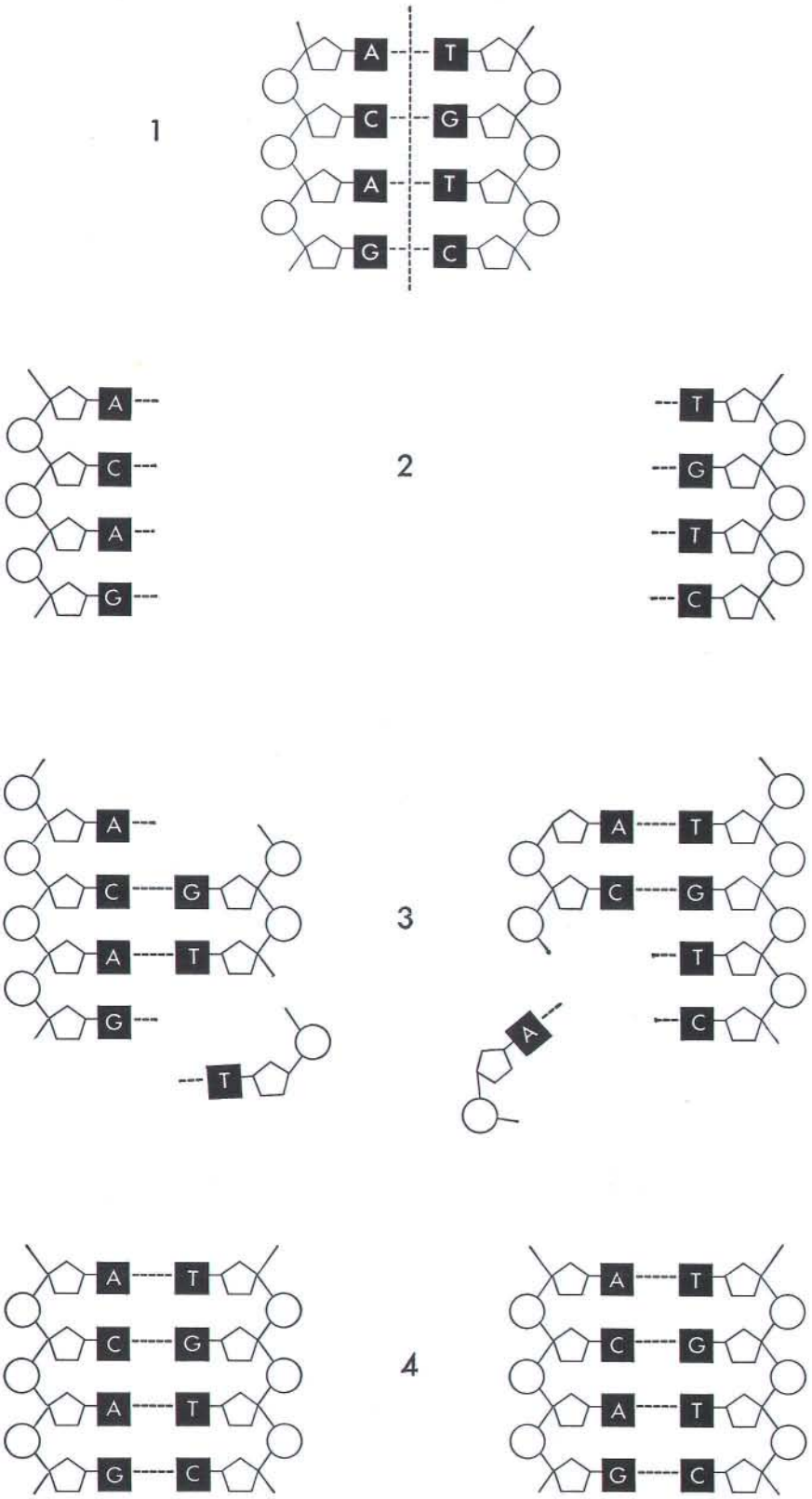
At the moment of writing, detailed interpretation of the X-ray photographs by Wilkins's group at King's College has not been completed, and until this has been done no structure can be considered proved. Nevertheless, there are certain features of the model that are so strongly supported by the experimental evidence that it is very likely they will be embodied in the final correct structure. For instance, measurements of the density and water content of the DNA fibers, taken with evidence showing that the fibers can be extended in length, strongly suggest that there are two chains in the structural unit of DNA. Again, recent X-ray pictures have shown clearly a most striking general pattern, which we can now recognize as the characteristic signature of a helical structure. In particular there are a large number of places where the diffracted intensity is zero or very small, and these occur exactly where one expects from a helix of this sort. Another feature one would expect is that the X-ray intensities should approach cylindrical symmetry, and it is now known that they do this. Recently



ONE LINKAGE of base to base across the pair of DNA chains is between adenine and thymine. For the structural arrangement proposed, the link of a large base with a small one is required to fit the two chains together.



ANOTHER LINKAGE is composed of guanine with cytosine. Assuming the existence of hydrogen bonds between the bases, these two pairings, and only these, will explain the actual configuration of the two DNA chains.



REPLICATION MECHANISM by which DNA might duplicate itself is depicted in this diagram. A helix of two DNA chains unwinds and separates (*1*). Two complementary chains of DNA (*2*) within the cell begin to attach DNA precursor units floating loosely (*3*). When the proper bases are joined, two new helixes will build up (*4*). The letters *A*, *T*, *C* and *G* represent the bases.

Wilkins and his co-workers have given a brilliant analysis of the details of the X-ray pattern of the crystalline form and have shown that they are consistent with a structure of this type, although in the crystalline form the bases are tilted away from the fiber axis instead of perpendicular, as in our model. Our construction was based on the paracrystalline form.

Many of the physical and chemical properties of DNA can now be understood in terms of this model. For example, the comparative stiffness of the structure explains rather naturally why DNA keeps a long, fiberlike shape in solution. The hydrogen bonds of the bases account for the behavior of DNA in response to changes in pH. Most striking of all is the fact that in every kind of DNA so far examined—and more than 40 have been analyzed—the amount of adenine is about equal to the amount of thymine and the guanine equal to the cytosine, while the cross-ratios (between, say, adenine and guanine) can vary considerably from species to species. This remarkable fact, first pointed out by Chargaff, is exactly what one would expect according to our model, which requires that every adenine be paired with a thymine and every guanine with a cytosine.

It may legitimately be asked whether the artificially prepared fibers of extracted DNA, on which our model is based, are really representative of intact DNA in the cell. There is every indication that they are. It is difficult to see how the very characteristic features of the model could be produced as artifacts by the extraction process. Moreover, Wilkins has shown that intact biological material, such as sperm heads and bacteriophages, gives X-ray patterns very similar to those of the extracted fibers.

The present position, therefore, is that in all likelihood this statement about DNA can safely be made: its structure consists of two helical chains wound around a common axis and held together by hydrogen bonds between specific pairs of bases.

Now the exciting thing about a model of this type is that it immediately suggests how the DNA might produce an exact copy of itself. The model consists of two parts, each of which is the complement of the other. Thus, either chain may act as a sort of mold on which a complementary chain can be synthesized. The two chains of a DNA, let us say, unwind

and separate. Each begins to build a new complement onto itself. When the process is completed, there are two pairs of chains where we had only one. Moreover, because of the specific pairing of the bases the sequence of the pairs of bases will have been duplicated exactly; in other words, the mold has not only assembled the building blocks but also has put them together in just the right order.

Let us imagine that we have a single helical chain of DNA and that floating around it inside the cell is a supply of precursors of the four sorts of building blocks needed to make a new chain. Unfortunately, we do not know the makeup of these precursor units; they may be, but probably are not, nucleotides, consisting of one phosphate, one sugar and one base. In any case, from time to time a loose unit will attach itself by its base to one of the bases of the single DNA chain. Another loose unit may attach itself to an adjoining base on the chain. Now if one or both of the two newly attached units is not the correct mate for the one it has joined on the chain, the two newcomers will be unable to link together, because they are not the right distance apart. One or both will soon drift away, to be replaced by other units. When, however, two adjacent newcomers are the correct partners for their opposite numbers on the chain, they will be in just the right position to be linked together and begin to form a new chain. Thus, only the unit with the proper base will gain a permanent hold at any given position, and eventually the right partners will fill in the vacancies all along the forming chain. While this is going on, the other single chain of the original pair also will be forming a new chain complementary to itself.

At the moment this idea must be regarded simply as a working hypothesis. Not only is there little direct evidence for it, but there are a number of obvious difficulties. For example, certain organisms contain small amounts of a fifth base, 5-methyl cytosine. So far as the model is concerned, 5-methyl cytosine fits just as well as cytosine, and it may turn out that it does not matter to the organism which is used, but this has yet to be shown.

A more fundamental difficulty is to explain how the two chains of DNA are unwound in the first place. There would have to be a lot of untwisting, for the total length of all the DNA in a single chromosome is something like four centimeters (400 million angstroms). This means that there must be more than 10 million turns in all, although the DNA may not be all in one piece.

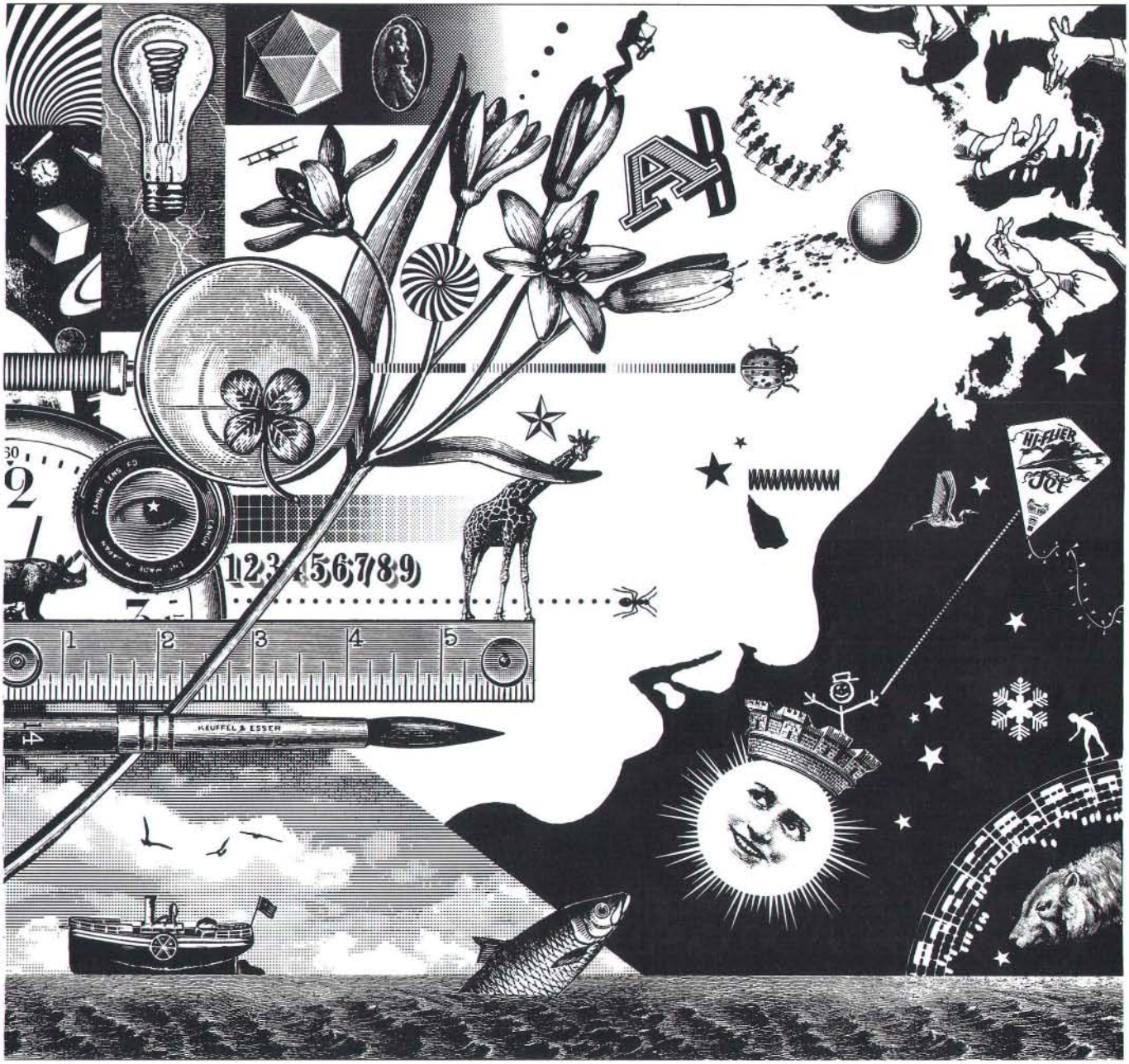
The duplicating process can be made to appear more plausible by assuming that the synthesis of the two new chains begins as soon as the two original chains start to unwind, so that only a short stretch of the chain is ever really single. In fact, we may postulate that it is the growth of the two new chains that unwinds the original pair. This is likely in terms of energy because, for every hydrogen bond that has to be broken, two new ones will be forming. Moreover, plausibility is added to the idea by the fact that the paired chain forms a rather stiff structure, so that the growing chain would tend to unwind the old pair.

The difficulty of untwisting the two chains is a topological one and is caused by the fact that they are intertwined. There would be no difficulty in "unwinding" a single helical chain, because there are so many single bonds in the chain about which rotation is possible. If in the twin structure one chain should break, the other one could easily spin around. This might relieve accumulated strain, and then the two ends of the broken chain, still being in close proximity, might be joined together again. There is even some evidence suggesting that in the process of extraction the chains of DNA may be broken in quite a number of places and that the structure nevertheless holds together by means of the hydrogen bonding, because there is never a break in both chains at the same level. Yet, in spite of these tentative suggestions, the difficulty of untwisting remains a formidable one.

There remains the fundamental puzzle as to how DNA exerts its hereditary influence. A genetic material must carry out two jobs: duplicate itself and control the development of the rest of the cell in a specific way. We have seen how it might do the first of these, but the structure gives no obvious clue concerning how it may carry out the second. We suspect that the sequence of the bases acts as a kind of genetic code. Such an arrangement can carry an enormous amount of information. If we imagine that the pairs of bases correspond to the dots and dashes of the Morse code, there is enough DNA in a single cell of the human body to encode about 1,000 large textbooks. What we want to know, however, is just how this is done in terms of atoms and molecules. In particular, what precisely is it a code for? As we have seen, the three key components of living matter—protein, RNA and DNA—are probably all based on the same general plan. Their backbones are regular, and the variety comes from the sequence of the side groups. It is therefore natural to suggest that the sequence of the bases of the DNA is in some way a code for the sequence of the amino acids in the polypeptide chains of the proteins that the cell must produce. The physicist George Gamow has recently suggested in a rather abstract way how this information might be transmitted. There are some difficulties with the scheme he has proposed, however, and so far he has not shown how the idea can be translated into precise molecular configurations.

What, then, one may reasonably ask, are the virtues of the proposed model, if any? The prime virtue is that the configuration suggested is not vague but can be described in terms acceptable to a chemist. The pairing of the bases can be described rather exactly. The precise positions of the atoms of the backbone are less certain, but they can be fixed within limits, and detailed studies of the raw data, now in progress at King's College, may narrow these limits considerably. Then the structure brings together two striking pieces of evidence which at first sight seem to be unrelated—the analytical data, showing the one-to-one ratios for adenine-thymine and guanine-cytosine, and the helical nature of the X-ray pattern. These can now be seen to be two facets of the same thing. Finally, is it not perhaps a remarkable coincidence, to say the least, to find in this key material a structure of exactly the type one would need to carry out a specific replication process: namely, one showing both variety and complementarity?

The model is also attractive in its simplicity. While it is obvious that whole chromosomes have a fairly complicated structure, it is not unreasonable to hope that the molecular basis underlying them may prove to be rather simple. If this is so, it may not prove too difficult to devise experiments to unravel it. It would of course help greatly if biochemists could discover the immediate precursors of DNA. If we knew the monomers from which nature makes DNA, RNA and protein, we might be able to carry out very spectacular experiments in the test tube. Be that as it may, we now have for the first time a well-defined model for DNA and for a possible replication process, and this in itself should make it easier to devise the crucial experiments.



Curiosity is the Frontier

Ever noticed how kids are always asking "Why?"

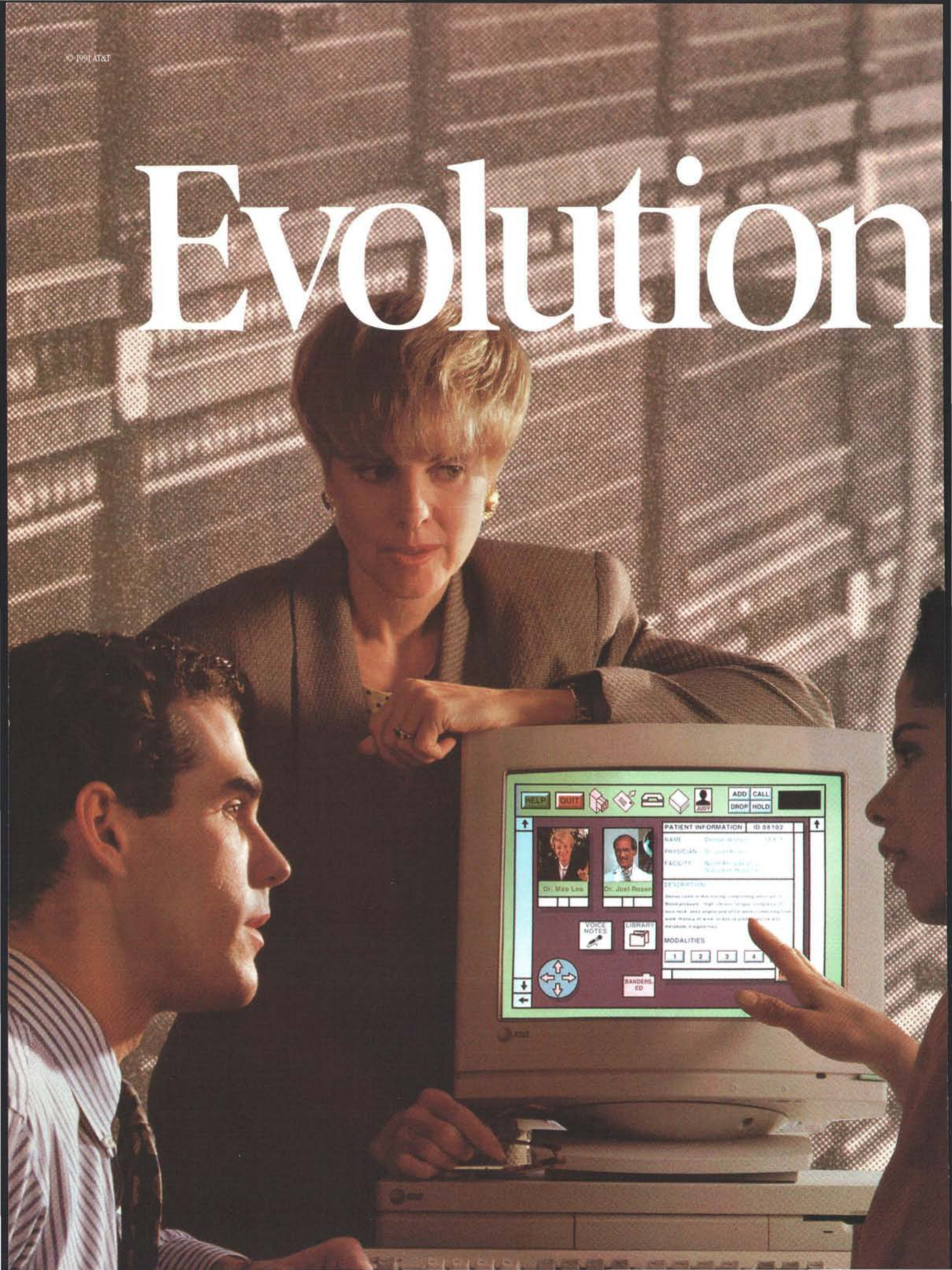
It's no surprise, really: All kids have a natural curiosity about the world around them. And they want all the answers.

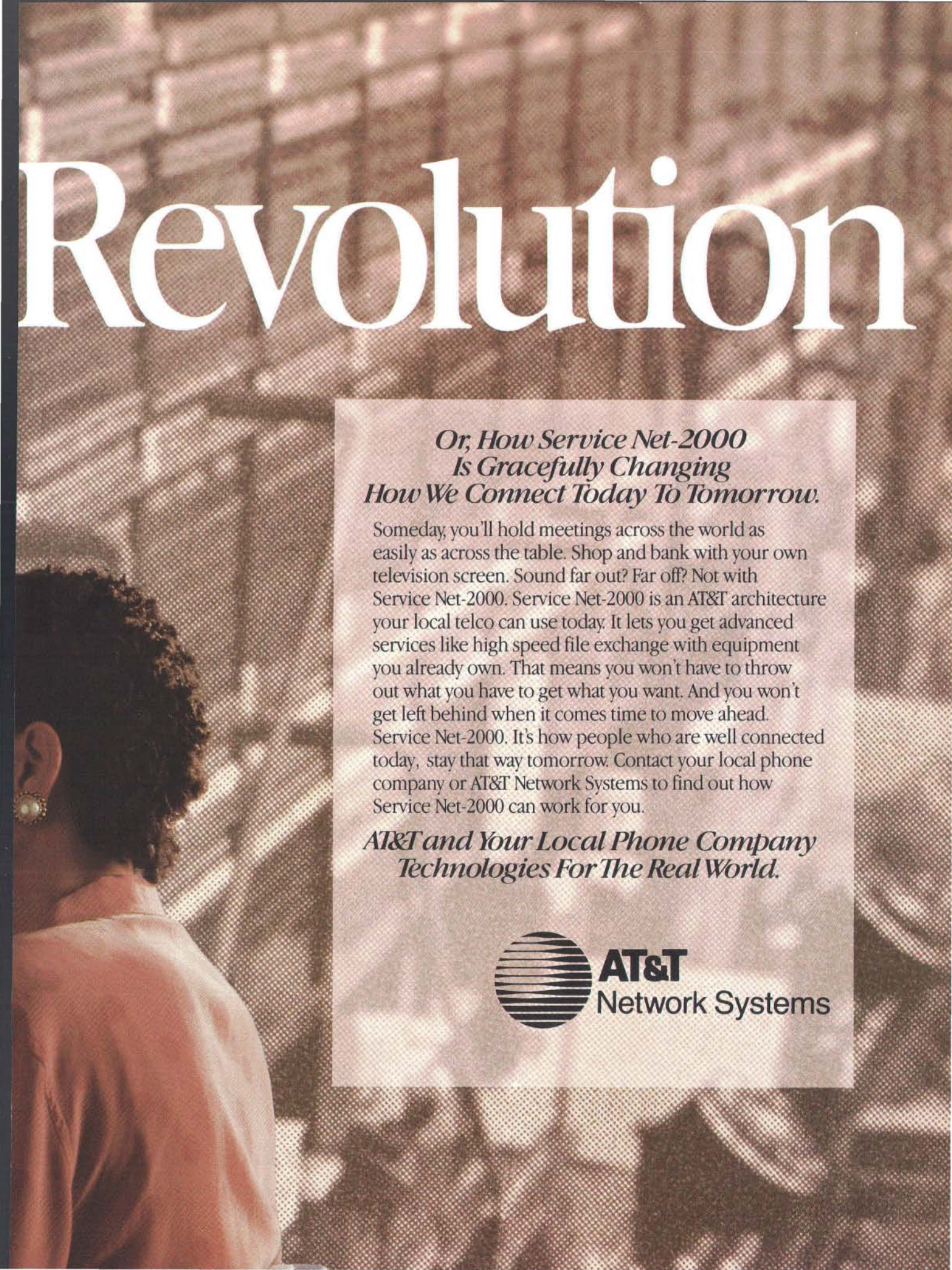
During National Science & Technology Week, April 21-27, we're encouraging you to ask questions together. Why do birds fly? Stars twinkle? Flowers open? And why does the sun disappear every night? Ask "Why" together—and watch your child's curiosity unfold!

NATIONAL SCIENCE & TECHNOLOGY WEEK

National Science Foundation
1800 G Street, N.W., Washington, DC 20550

Evolution




Revolution
Or, How Service Net-2000
Is Gracefully Changing
How We Connect Today To Tomorrow.
Someday, you'll hold meetings across the world as easily as across the table. Shop and bank with your own television screen. Sound far out? Far off? Not with Service Net-2000. Service Net-2000 is an AT&T architecture your local telco can use today. It lets you get advanced services like high speed file exchange with equipment you already own. That means you won't have to throw out what you have to get what you want. And you won't get left behind when it comes time to move ahead. Service Net-2000. It's how people who are well connected today, stay that way tomorrow. Contact your local phone company or AT&T Network Systems to find out how Service Net-2000 can work for you.
AT&T and Your Local Phone Company
Technologies For The Real World.
AT&T
Network Systems

The Molecules of Life

The advances of recent years have transformed biology from a descriptive science to a means of understanding and manipulating the molecular mechanisms of life

by Robert A. Weinberg

Biology in 1985 is dramatically different from its antecedents only 10 years ago. New investigative techniques have made commonplace many experiments that were previously far beyond the reach of even the cleverest experimental biologist. The new molecular biology has done much more than expand the repertoire of laboratory techniques. It has, with remarkable rapidity, established a biotechnology industry. More important, it has changed the ways people think about living things, from bacteria to human beings.

Biology has traditionally been a descriptive science. The multitude of living organisms were catalogued, their traits enumerated and their structures examined on a gross or a microscopic level. In thus describing organismic traits, or phenotypes, biologists confronted only the consequences of biological processes, not the causative forces. An experimenter could watch a muscle contract or an embryo develop, but such observation alone could not provide the clues that were needed for any real understanding of underlying mechanisms.

The ability to observe was greatly extended by the development of microscopic techniques that made it possible to visualize cells and subcellular organelles. Electron microscopy pushed the limits of visualization even further: the fine structure of cells could be resolved with great precision. This advance led to the uncovering of still more structures and phenomena whose causative mechanisms remained unexplained. The explanations clearly lay with elements even smaller than the cellular components observed by microscopists.

It became apparent that the ultimate causal mechanisms behind many biological phenomena depend on the functioning of specific molecules inside and outside the cell. Hence, investigators now think about biological systems in terms of their molecular components. Indeed, the current assumption is that to describe biological phenomena is far less interesting than to elucidate the molecular mechanisms underlying them. Molecular biologists manipulate things they will never see. Yet they work with a certainty that the invisible, submicroscopic agents they study can explain, at one essential level, the complexity of life.

The newly gained ability to describe and manipulate molecules means the biologist is no longer confined to studying life as the end product of two billion and more years of evolution. The new technology has made it possible to change critical elements of the biological blueprint at will and in so doing to create versions of life that were never anticipated by natural evolution. In the long run this may prove to be the most radical change deriving from the power to manipulate biological molecules.

Among the many kinds of biological molecules in the living cell, three have attracted the greatest attention: protein, RNA and DNA. They are macromolecules, that is, large molecules that are linear polymers built up from simple subunits, or monomers. It was the proteins that attracted the lion's share of attention until 20 years ago. The reason, in retrospect, is clear. Certain specialized tissues accumulate large amounts of only one kind of protein. Red blood cells have almost pure hemoglobin, cartilage consists largely of collagen and hair is largely keratin. Biochemists studied such proteins first because they could be isolated in pure form, purity being a prerequisite to further study.

As an array of sophisticated biochemical techniques emerged, it became possible also to purify those proteins found only in trace amounts within the complex chemical soup of a living cell. Biochemists could now concentrate on proteins that function as enzymes, catalyzing the several thousand biochemical reactions that in the aggregate constitute the metabolism of living cells. This work went well, because many of the reactions could be easily reconstructed in a test tube containing the proper mixture of reactants and catalyzing enzymes.

Yet in the past quarter of a century proteins have been gradually upstaged as objects of attention by the other macromolecules, first by RNA and more recently by DNA. There were two important reasons. The first one stems ironically from the great successes of protein biochemistry, which produced

ROBERT A. WEINBERG is professor of biology at the Massachusetts Institute of Technology and the Whitehead Institute for Biomedical Research. His B.A. (1964) and Ph.D. (1969) are both from M.I.T. He did postdoctoral research with Ernest Winocur at the Weizmann Institute of Science in Israel and with Renato Dulbecco at the Salk Institute for Biological Studies. In 1972 he returned to M.I.T., and the following year he was made a member of the faculty at the Center for Cancer Research there. In 1982 Weinberg became professor and joined the Whitehead Institute.

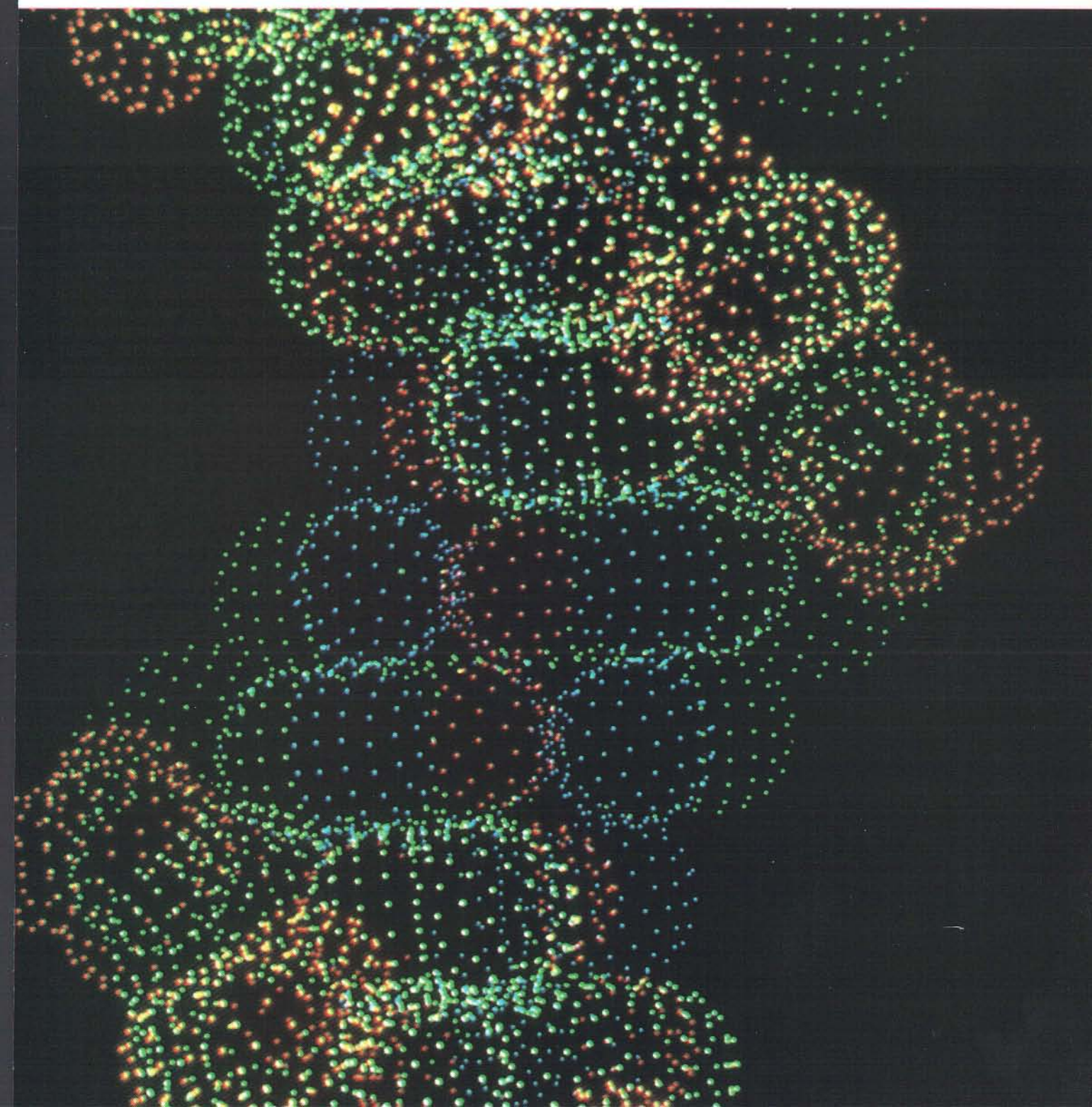
DOUBLE HELIX OF DNA, the molecule that is the repository of genetic information and so may be considered the fundamental molecule of life, is seen from the side in the computer-generated image on the opposite page. The spheres represent individual atoms that make up the molecule: oxygen is red, nitrogen is blue, carbon is green and phosphorus is yellow. The diagonal regions of the image delineate the sugar-phosphate backbone of the ladderlike helix; the horizontal elements, made up of nitrogen, carbon and oxygen atoms, are the base pairs that cross-link the two strands of the helix. The computer image was generated by the Computer Graphics Laboratory of the University of California at San Francisco.

Reprinted from the October 1985 issue

an avalanche of data on thousands of proteins and biochemical reactions. It soon became apparent that further study of such individual trees gave little hope of understanding the entire forest. What was responsible for organizing and orchestrating this complex array of structures and processes? The answer lay not with the proteins but with the study of genetics and of the nucleic acids that carry genetic information.

The other reason nucleic acids, particularly DNA, have taken center stage is the advent of recombinant DNA technology. In the course of the past decade biologists have learned to manipulate DNA in ways that (at least currently) are impossible for the protein chemist. DNA can be cut apart, modified and reassembled; it can be amplified to many copies; perhaps most telling, with DNA one can generate RNA and then protein molecules of wanted size and constitution. The central experimental maneuver in these manipulations is the cloning of genes, and it is cloning, more than any other single factor, that has changed the face of biology.

The groundwork for the cloning of genes was laid in 1953, when the double-helical structure of DNA was perceived by James Watson and Francis Crick. A strand of DNA is a chain of nucleotides, each containing one of four bases: adenine (*A*), guanine (*G*), thymine (*T*) and cytosine (*C*). An *A* on one strand of the double helix pairs with a *T* on the other strand, and *G* pairs with *C*, so that the two strands



are complementary. The sequence of bases specifies the order in which amino acids are assembled to form proteins. When the information in a gene is read out (expressed), its base sequence is copied (transcribed) into a strand of RNA. This messenger RNA (mRNA) serves as a template for the synthesis of protein: its base sequence is translated into the amino acid sequence of one protein or another.

The encoding of proteins is only a small part of DNA's function and hence of its information content. To learn this and other simple facts, it was necessary to first learn about the overall organization of DNA sequences and how the functional units of DNA—the individual genes—interact with one another in the total genetic repertoire of the organism, which is called its genome.

The genome of complex organisms resisted analysis until recently. Analysis of the gross biochemical properties of cellular DNA gave little hope of understanding the subtleties of genetic organization. The DNA content of even a bacterial cell is very large; the much larger genome of a mammalian cell carries some 2.5 billion base pairs of information arrayed along its chromosomal DNA. The base sequences are arranged in discrete compartments of information, that is, the individual genes. There are between 50,000 and 100,000 genes in the genome of a mammal; each one is presumably responsible for specifying the structure of a particular gene product, usually a protein. Interest was therefore focused on studying individual genes, but that undertaking was doomed until recently by an inability to study single genes in isolation. In the absence of effective techniques of enrichment and isolation, individual cellular genes were abstractions. Their existence was suggested by genetic analysis, but their physical substance remained inaccessible to direct biochemical analysis.

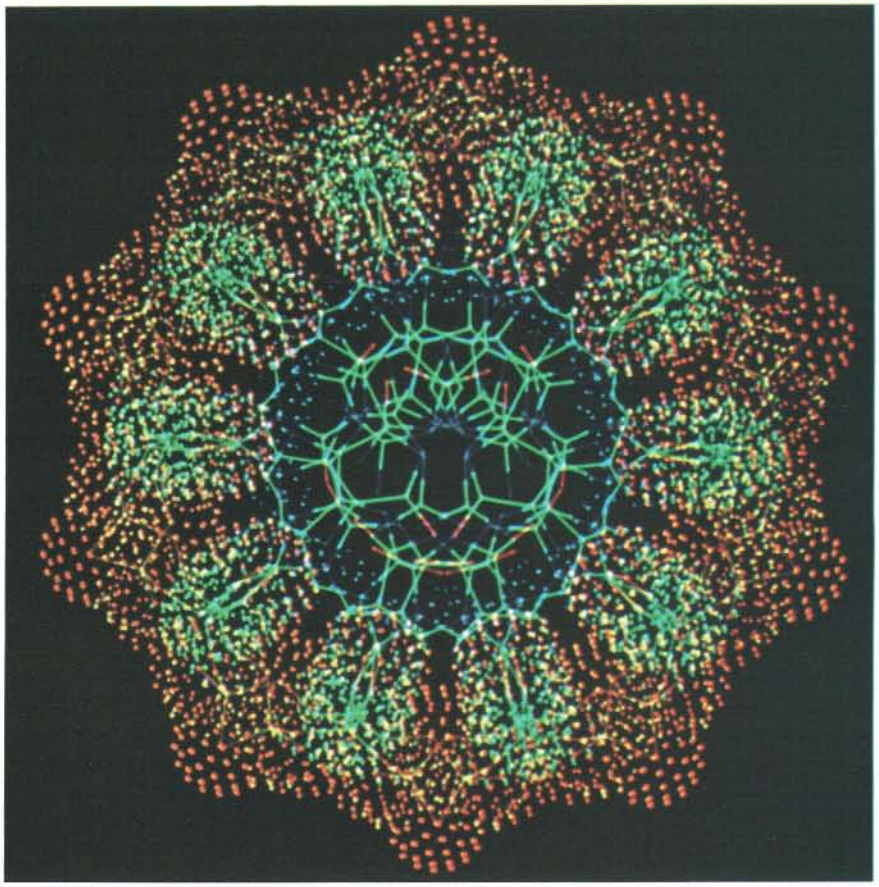
A partial solution to this quandary came from the study of viruses. Their genome is very small compared with that of a cell, and yet their genes are similar to those of the cells they infect. The DNA genome of one much studied animal virus, the SV40 virus of monkeys, has only 5,243 base pairs, in which are nested five genes. The analysis of an individual gene was therefore not confounded by a large excess of unrelated sequences. Moreover, the viral genome multiplies to several hundred thousand identical copies within an infected cell, and it was not hard to separate the viral DNA from the cellular DNA.

Once purified, the relatively simple viral DNA made it possible to study aspects of gene structure, the transcription and processing of messenger RNA and the synthesis of proteins that had previously been beyond reach. Still unresolved were the detailed structure and the base sequence of even the viral genome, whose 5,000-odd base pairs represented a daunting challenge to traditional biochemists trained to take polymers apart one unit at a time. Then, in the mid-1970s, two revolutionary techniques became widely available that radically simplified the analysis of DNA structure.

The first of the techniques stemmed from the discovery of DNA-cleaving enzymes called restriction endonucleases. These enzymes, extracted from bacteria, cut a DNA molecule only at specific sequences that occur here and there along the DNA double helix. The much used endonuclease *Eco*RI, for example, cuts wherever it encounters the sequence *GAATTC; Sma*I cuts at *CCCGGG*, and so on. The sequences forming recognition sites occur with a certain statistical probability in any stretch of DNA.

Restriction enzymes have become powerful tools for experimenters. They establish convenient, fixed landmarks along the otherwise featureless terrain of the DNA molecule. They allow one to reduce a very long DNA molecule into a set of discrete fragments, each of them ranging in length from several hundred to several thousand bases. The fragments can be separated from one another on the basis of their size by gel electrophoresis. Each fragment can then be subjected individually to further analysis.

The other technical revolution had to do with the sequencing of DNA. Several procedures were invented by which the entire base sequence of a segment generated by restriction-enzyme cleavage can be determined with remarkable rapidity. These methods made it possible, for example, to establish the entire nucleotide sequence of the SV40 genome by 1978. Because the genetic code for translating a DNA sequence



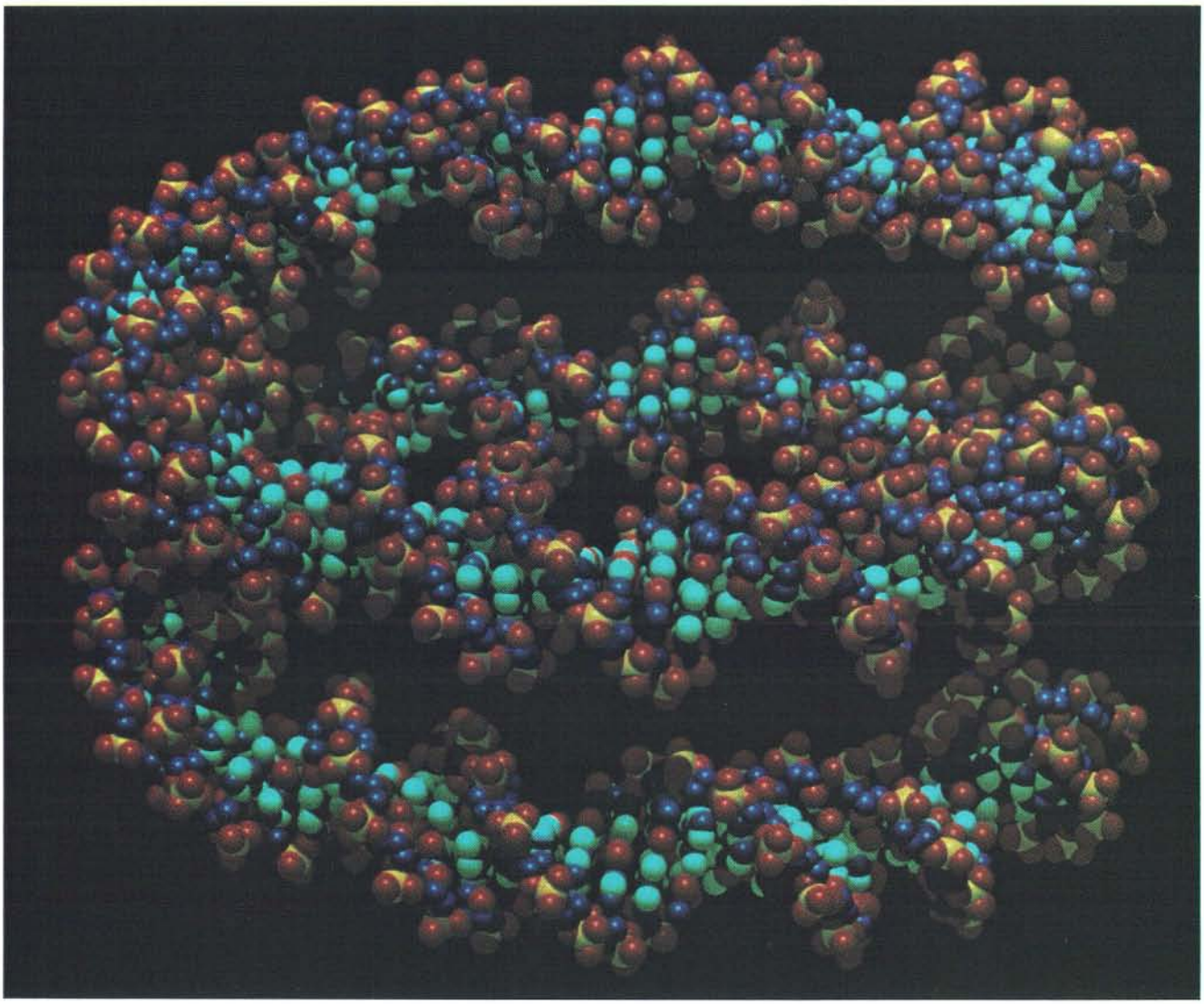
MOLECULAR ROSE WINDOW is a view along the axis of the *B* DNA double helix. In this image, also made by the Computer Graphics Laboratory, 10 consecutive nucleotide pairs along the helix are collapsed into a plane; the tenfold symmetry results from the fact that there are 10 component nucleotides per turn of the helix. The surfaces of the sugar and phosphate groups of the backbones are delineated by dots representing atoms: carbon (*green*), oxygen (*red*) and phosphorus (*yellow*). In the center the dots are absent, and as a result the skeletal structure of the bases, largely nitrogen (*blue*) and carbon, is left exposed.

into an amino acid sequence was already known, the base sequences in certain regions of the genome could be translated into amino acid sequences. In this way, the structure of SV40's proteins could be deduced from its DNA structure. Previously, protein structure had been determined by the painstakingly slow biochemical analysis of individually isolated proteins; now the rapid sequencing of DNA could determine protein sequences in a fraction of the time. DNA sequencing also revealed other regions of the SV40 genome that are involved not in encoding proteins but in regulating the expression of genes and in the replication of DNA.

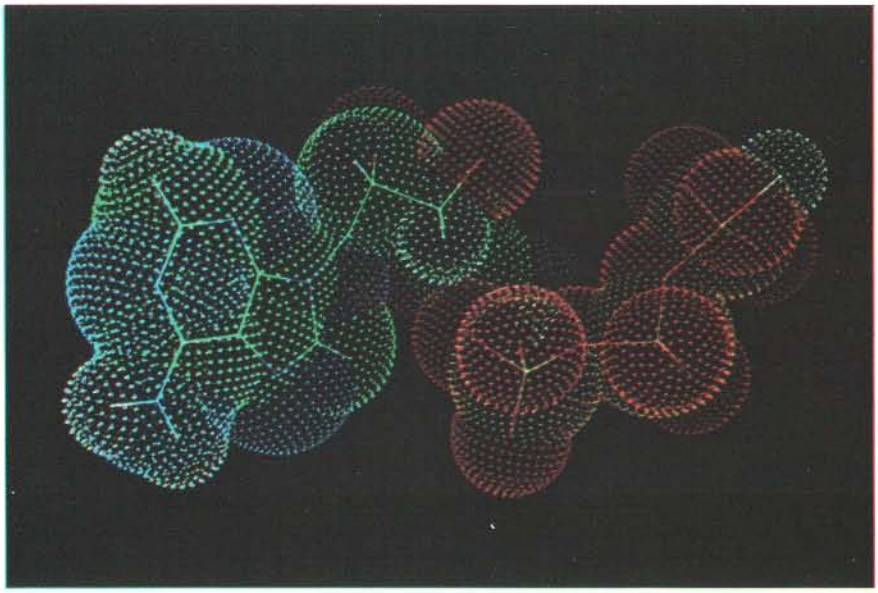
Further progress depended on procedures for isolating individual cellular genes. Here success came from studies of bacterial genetics that were initiated in the early 1970s. The procedures for gene isolation that grew out of this work are based ultmately on the similarity of molecular organization in all organisms, from bacteria through mammals. As a result, bacterial and mammalian DNAs are structurally compatible; DNA segments from one life-form can readily be blended with the DNA of another form.

This similarity in DNA structure extends to many of the bacteriophages (viruses that infect bacteria) and plasmids (small DNA circles that parasitize bacteria). Phages inject their DNA into a bacterium, cause it to be replicated many times over, package the newly replicated DNA in viral-protein coats and kill the bacterium; the progeny phages released from the cell go on to infect other susceptible hosts. The plasmids are even simpler, and they have a more symbiotic relation with the bacterial cells in which they grow. They may carry genes that confer advantages on their host cell, such as resistance to an antibiotic. The host cell in turn allows the plasmid DNA to be replicated to a limited extent in the cell, thereby ensuring the continued presence of the plasmid in the daughter bacteria arising when a parent bacterium divides.

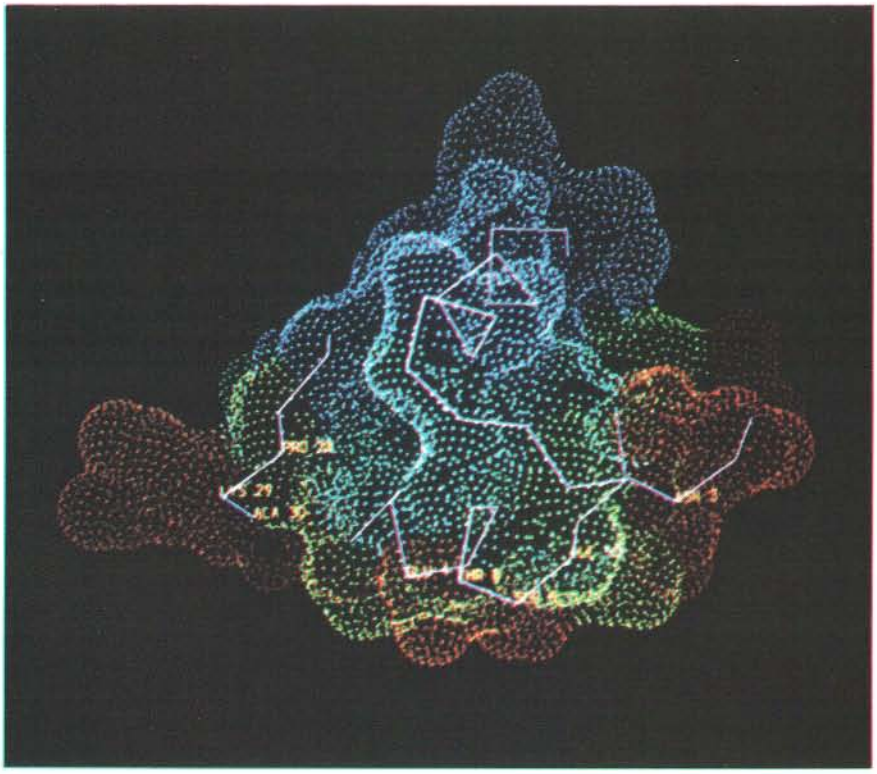
Some phage and plasmid DNAs are (like SV40 DNA) small in size, ranging in complexity from several thousand to 50,000 bases. Because of their small size, they can be manipulated and restructured by a variety of recently developed tools. The molecules are easily isolated, unbroken and in large amounts. They can be cut at a number of defined sites with restriction enzymes, and the resulting fragments can be rejoined with one another or joined to foreign DNA segments to reconsti-



DNA SUPERHELIX, in which the double helix is itself wound into a coil, is depicted in a computer model made by Nelson L. Max of the Lawrence Livermore National Laboratory. This is thought to be the form in which DNA is actually packed into chromosomes in the cell nucleus, with two turns of the superhelix being wound around a complex of histone proteins. The model is based on data from Joel L. Sussman and Edward N. Trifonov of the Weizmann Institute of Science.



ATP (adenosine triphosphate) is the molecule that provides free energy for many biochemical reactions, including those required for the polymerization of DNA, RNA and protein. ATP is modeled in this image made by the Computer Graphics Laboratory. It is a nucleotide consisting of the base adenine (*left*), a ribose sugar and three phosphate groups (*right*). Energy is acquired when a third phosphate is added to adenosine diphosphate by the oxidation of fuel molecules or, in plants, by photosynthesis; energy is liberated when ATP is broken down, freeing this third phosphate group. The skeletal structure of the molecule is indicated by the lines; the dots delineate the effective surfaces of the constituent atoms.



INSULIN MOLECULE, a hormone that has multiple functions, is depicted in a computer-generated model. It was developed by Elizabeth D. Getzoff, J. A. Tainer and Arthur J. Olson of the Research Institute of Scripps Clinic. Insulin is a small protein hormone made up of two short folded chains of amino acids. The lines trace the backbone of the two amino acid chains; the dots delineate the solvent-accessible surface. Coloring reflects the relative mobility of constituent atoms: the atoms shown in red and orange are the ones most subject to excursion from their mean position in a crystal of insulin, and those in green and blue are the least mobile.

tute the original molecule or make a hybrid molecule. The rejoining is done with readily available enzymes of bacterial origin known as DNA ligases, which recognize the ends of DNA molecules and fuse them without leaving any trace of the joining.

A hybrid DNA made of a plasmid fused with foreign (say, mammalian) genetic material can replicate when it is introduced into a bacterial cell. This means the plasmid genome can serve as a "vector" for establishing and amplifying the foreign DNA in bacteria. A phage vector functions similarly, and it can serve as well to convey the foreign DNA from one bacterium to another. When the vector DNA is copied in the course of replication, the inserted foreign DNA is copied, too.

The process of cloning begins with whole cellular DNA of an organism such as a mammal. The DNA is cleaved into fragments of a size (from about 1,000 to about 30,000 bases) that can be accommodated by the carrying capacity of one or another vector. A complex genome such as the human one can be broken down into a few hundred thousand DNA fragments. Each fragment can be separately inserted into a vector DNA molecule. The process does not require painstaking molecule-by-molecule assembly by a patient technician. Instead millions of inserted and vector DNA molecules are mixed together, and the process is completed in minutes by the addition of DNA ligase. If the resulting collection of hybrid molecules is large enough, any single gene of interest will surely be found embedded in one or another of the DNA segments linked to the vector molecules.

Each of these hybrid molecules, part vector and part inserted mammalian DNA, can now be introduced into a bacterial cell, where they are replicated many times over; each hybrid molecule spawns a separate progeny population, all of whose members are identical with the founder. Such a population is often called a clone to reflect its descent from a common ancestor and the identity of all its members.

The term "clone" has acquired another meaning. It is applied specifically to the bits of inserted foreign DNA in the hybrid molecules of the population. Each inserted segment originally resided in the DNA of a complex genome amid millions of other DNA segments of comparable size and complexity. When the manipulations described above are completed, the same segment is present in pure form within the confines of the particular clonal population, contaminated only by the

associated vector DNA. The inserted DNA segment has been isolated from its previous surroundings and selectively amplified: it has been cloned, and so the purified DNA insert itself is often called a clone.

The process of cloning requires one further step, which is usually the most challenging of all. The insertion and amplification process has given rise to hundreds of thousands of different clonal populations, each descended from a single hybrid DNA molecule. If the initial hybrids were properly diluted before being amplified, each descendant clonal population is physically separated from other populations carrying different inserted DNAs. Having established this large array (a "library") of distinct clonal populations, the experimenter is now faced with having to identify the one or several populations carrying the inserted DNA of interest.

Identification can be simple if a related gene or DNA segment has been cloned before. The previously cloned DNA can be labeled with a radioactive isotope; under appropriate conditions the radioactive DNA will preferentially stick to the clone of interest (because complementary DNA strands "hybridize" by base pairing) and thus identify it. The most interesting cloning is done, however, to isolate genes that have never been cloned before, even in related form. A variety of clever strategies have been developed to address this challenge. The goal is to develop a specific probe with which to scan a library of clones and identify the clones of interest.

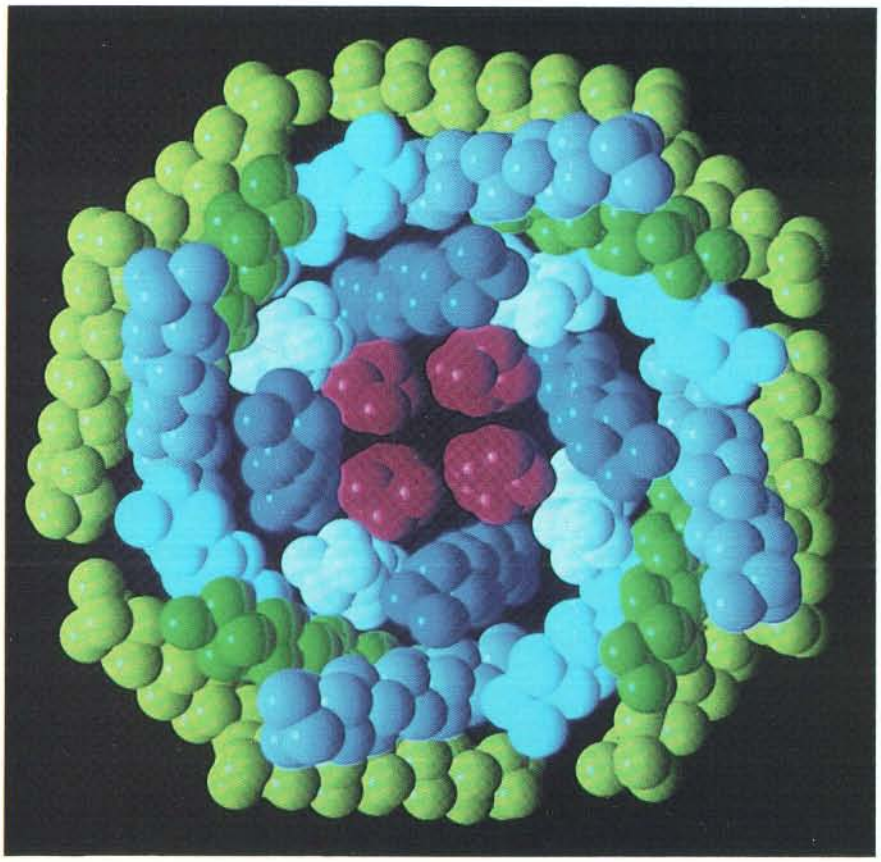
One strategy for probe development depends on the fact that some proteins are expressed at a high level in specialized cells. In the precursors of red blood cells, for example, globin (the protein component of hemoglobin) is synthesized in much larger amounts than any other protein. The mRNA that directs its synthesis is also present in large quantity, and there are ways to isolate it readily from other mRNAs in the same cell. The isolated mRNA, or a DNA copy of it, can serve as a probe that will hybridize with the corresponding gene sequence in a genomic library. Sophisticated versions of this strategy allow the mRNA encoding a protein of interest to be isolated selectively from a thousandfold excess of other mRNA molecules present in the same cell.

Often the protein whose gene is sought is rare, so that its mRNA cannot easily be isolated. In such cases, a small amount of the protein is purified and the amino acid sequence of some part of it is determined. Knowing the genetic code, one can back-translate the amino acid sequence to learn what DNA base sequences are likely to be present in the gene encoding the protein. Small pieces of DNA corresponding to these derived base sequences can be synthesized by assembling off-the-shelf nucleotides. These man-made gene fragments serve as probes for identifying the clone of interest.

Yet another strategy that begins with a protein depends on antibodies directed against the protein whose gene one seeks to clone. Bacteria infected by a phage carrying the wanted gene synthesize small quantities of the protein, and so a phage library can be screened by the proper antibody, which binds to the protein and thereby identifies the gene-carrying clone.

With these and other experimental strategies available, the technology is now at hand to clone any gene whose protein product is known and can be isolated in even a small amount. Given sufficient interest, any of the genes encoding the many hundreds of enzymes that have been studied by biochemists can be isolated. The genes for the important structural proteins of the cell, including those that determine cellular architecture, have been cloned. Other genes encoding such intercellular messengers as insulin, interferon, the interleukins and a number of growth factors have been isolated. Indeed, genes are being cloned and their sequences deciphered faster than the new data can be fully interpreted. Most of the sequences are now being stored in computer banks; perhaps future generations of biologists, aided by new analytical procedures, will be able to interpret them fully.

The genes specifying known proteins account for only a small part of a complex organism's total genetic repertoire. Most of the remaining genes probably encode proteins, too, but so far their existence is implied only by the effects



SODIUM CHANNEL, a large protein molecule embedded in the membrane of nerve cells, has been modeled by H. Robert Guy of the National Cancer Institute; this image of one model was computer generated by Richard J. Feldmann of the National Institutes of Health. The protein admits sodium ions to the neuron, thereby supporting the action potential, the voltage pulse that ultimately triggers the release of neurotransmitter. The protein has four homologous domains; each domain includes eight distinct protein substructures. Similarly colored groups of spheres represent the four homologous versions of each substructure. (Two substructures in each domain are very similar and are both shown in pale green.)

they exert on cellular and organismic structure and function. Some of them specify biochemical conversions in the cell, others govern complex developmental processes that create shape and form in a developing embryo and still others may specify behavioral attributes of an organism. Such genes remain elusive because the means of identifying them in genomic libraries are limited.

The flow of genes from genome to gene library makes more things possible than the detailed description of DNA and protein structure. Once cloned, a gene can be inserted into a foreign cell, which can be forced to express it. The cell then synthesizes the protein the gene specified in its original home.

The gene to be expressed is excised from the vector in which it was cloned and subjected to important modifications. The modifications are necessary because a mammalian gene carries regulatory sequences that promote its transcription into mRNA in its home cell, not in a bacterial cell. These need to be replaced by bacterial regulatory sequences. The modified gene is then introduced into an "expression vector": a plasmid designed to facilitate the expression of the gene in a foreign cellular environment. The mammalian gene (or a similarly engineered plant gene) carrying bacterial regulatory sequences is then introduced into a selected foreign host, usually a bacterial or yeast cell.

A protein that is synthesized only in limited amounts in its normal host can be produced in large quantity when its gene is redesigned for high-level expression in bacteria or yeast. This can confer great economic advantage and represents a cornerstone of the biotechnology industry. Microorganisms bearing cloned genes can be grown quite cheaply in large volume in fermentation chambers, leading to an enormous scale-up in protein production. Among the products currently being manufactured or being considered for manufacture are insulin, interferon (for combating infections and perhaps tumor growth), urokinase and plasminogen activator (for dissolving blood clots), rennin (for making cheese from milk), tumor necrosis factor (for possible cancer therapy), the enzyme cellulase (to make sugar from plant cellulose) and viral peptide antigens (for creating novel and safe vaccines).

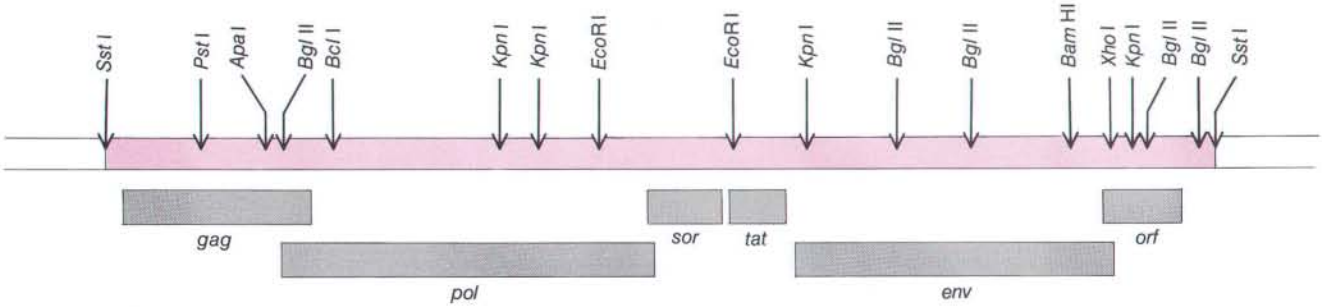
In a different version of gene insertion, mammalian genes that have been cloned in bacteria are introduced into mammalian cells grown in culture rather than into microorganisms. Although cultured mammalian cells cannot be grown economically in the large numbers that characterize a bacterial or yeast culture, they have the advantage of being able to make minor but significant modifications to proteins encoded by mammalian genes. For example, certain mammalian proteins function better when sugar and lipid side chains have been attached to their amino acid backbone. The addition takes place routinely in mammalian cells but not in bacteria.

Cloned genes can now be inserted not only into microorganisms or cultured mammalian cells but also into the genome of an intact multicellular plant or animal. Here the motives are quite different from those governing the genetic engineering of unicellular microorganisms to achieve large-scale production of desirable gene products. Plants and animals can be modified genetically in an effort to alter such organismic traits as growth rate, disease resistance and ability to adapt to novel environments.

Gene insertion into a multicellular organism is a quite different project from gene transfer into a single cultured cell. The introduction of a cloned gene into most types of cells in a plant or an animal can alter the behavior of only those few cells that acquire the gene. Obviously it is of far greater interest to imprint the change on an entire organism and on the organism's descendants. That calls for gene insertion specifically into germ cells (sperm or eggs), which transmit genetic information from parent to offspring.

Techniques are indeed now available for achieving germ-line insertion into mammals, flies and certain plants. It is done either by direct physical injection of a cloned gene into the early embryo or by the use of a viral vector to carry the gene into the cells of an embryo. Again the resulting animal (or plant) carries the inserted gene in only some of its cells, but now one can hope the gene is in some of its germ cells. The presence of the gene in germ cells may allow some of the organism's offspring to inherit the inserted gene along with other parental genes, so that it will be present in all their cells. Thus incorporated into the germ line of the progeny organisms, the gene is passed on to, and affects, the descendants of those organisms.

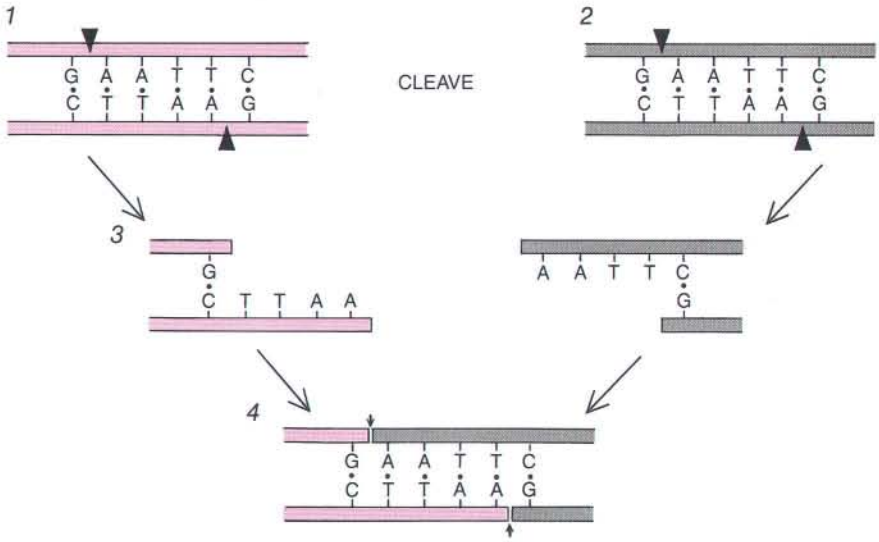
Techniques for germ-line insertion are still limited in important ways, and they may forever be. They cannot direct the foreign DNA to insert itself (to "integrate") into a particular chromosomal site; the locus of insertion is random. They cannot supplant an existing gene in the organism by knocking it out; rather they merely add incrementally to the existing genome. More-



RESTRICTION-ENZYME MAP of the genome of the virus HTLV-III, the AIDS agent, was developed in the laboratory of Robert C. Gallo of the National Cancer Institute. Such a map is the primary means by which molecular biologists depict the organization of a stretch of DNA. The DNA is cleaved with a restriction endonuclease, an enzyme that cuts DNA at specific sites. The size of the fragments is known from the distance they travel through an electrophoresis gel. Individual fragments can be cloned and sequenced. Cleaving a genome with a number of different restriction enzymes provides additional mileposts. This relatively simple map of HTLV-III shows the sites where several restriction enzymes cleave the HTLV-III DNA (*top*) and the locations of the various genes (*bottom*). For example, the surface antigen of the virus is encoded by the *env* (for envelope) gene, and the enzyme (reverse transcriptase) that copies the RNA of the virus into DNA is encoded by *pol* (polymerase). The total length of this DNA from the AIDS virus is 9.3 kilobases (thousands of bases).

over, inserted genes do not always function precisely like resident genes, which are turned on and off at appropriate times in development.

Germ-line insertion is nonetheless powerful. Mice have been developed that carry and transmit to their offspring the genes for extra growth hormones. Giant mice (half again as large as normal) ensue; cattle with altered growth properties will soon follow. Flies have been developed that carry a variety of inserted genes, leading to novel insights into the molecular biology of fly development. Plants are being developed that carry genes conferring resistance to herbicides. As gene-insertion techniques are improved and as additional genes are cloned, the possibilities for altering organismic traits will expand enormously. The molecular biologist will no longer confront living forms as the finished products of evolution but will be an active participant in initiating organismic change.



CLONING IS FACILITATED by "sticky ends" generated by some restriction enzymes. The enzyme *Eco*RI, for example, makes a staggered cut in the sequence *GAATTC*. When genomic DNA (*1*) and a vector DNA (*2*) are cut with *Eco*RI, the ends of the resulting fragments have single-strand projections of complementary bases (*3*). When the fragments are mixed, hydrogen bonds (*dots*) form between those bases, reversibly joining genomic and vector DNAs (*4*). The joint is sealed irreversibly with the enzyme DNA ligase (*arrows*).

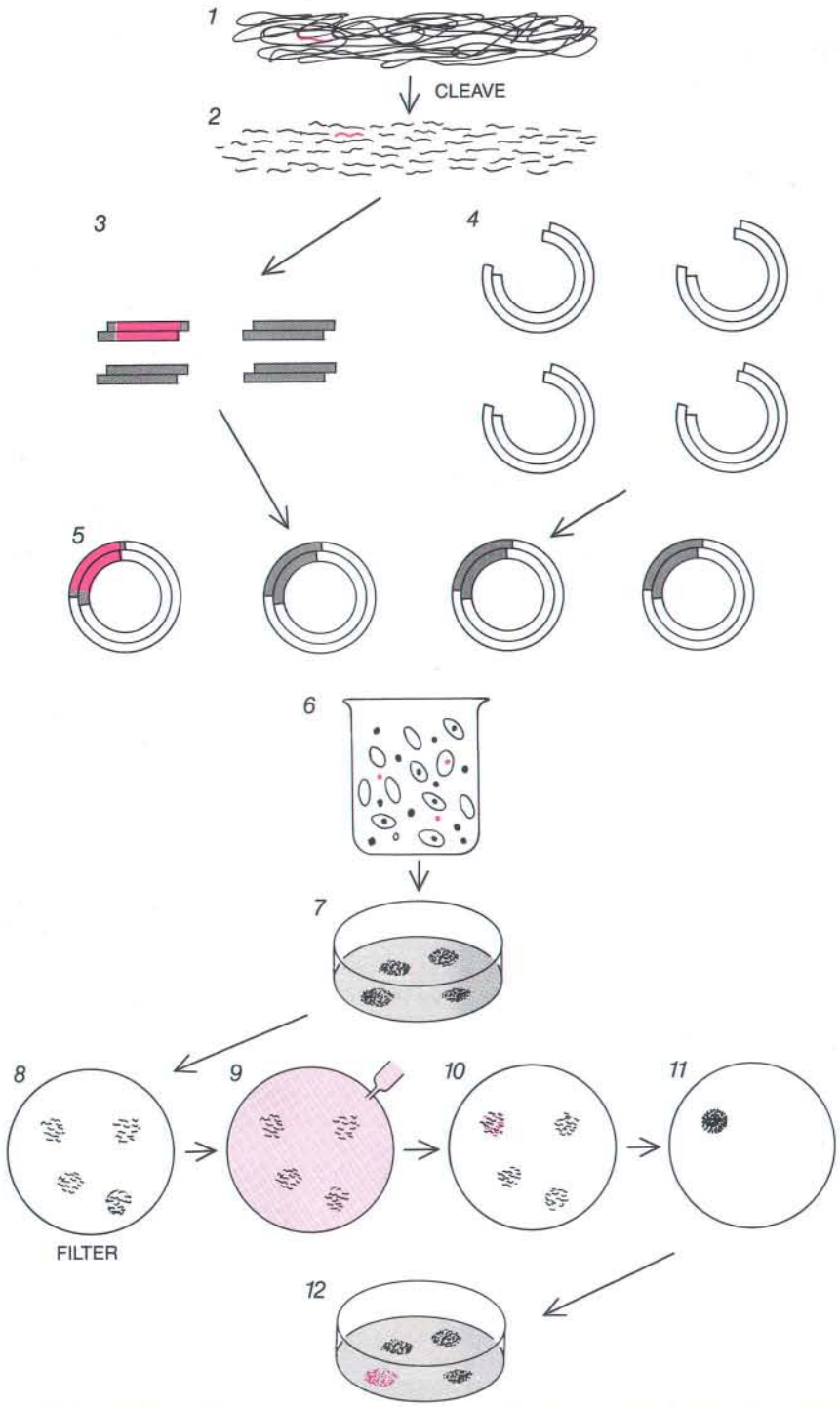
For experimental biologists, cloning and its associated methods have attained and will retain a preeminent role. Cloning makes it possible to analyze a biological system at three levels. First, the genes relevant to a particular biological problem can be isolated, the sequence of the DNA can be elucidated and the functioning and regulation of the DNA can be revealed. Second, once the DNA of a gene has been cloned, the RNA transcribed from it can be produced in large amounts for study. The RNA can act in many ways to modulate the expression of genes; RNA structure and processing are central to a full understanding of gene function. Third, what is perhaps the greatest advantage of cloning stems from the analysis of the proteins encoded by a gene. How do various proteins act to elicit myriad responses in the cell? Proteins that formerly were available for study in minute amounts can now be made in great quantities once their gene has been isolated. In sum, all the major macromolecular components of a biological system can now be made available in large amounts, in pure form.

Equally important is a newly gained ability to perturb biological systems. Genes and their encoded proteins can be redesigned so that new functions can be imparted to DNA and proteins. The relations among the interacting components of a biological system can be altered to generate novel and often revealing behavior by the system as a whole. The redesigning of genes is accomplished by changing DNA sequences through what is termed site-directed mutagenesis. This may involve the replacement of one restriction-enzyme fragment with another in the midst of a cloned gene. Alternatively, chemically synthesized DNA segments may be stitched into a gene, replacing or adding to existing sequence information. Single nucleotides can be substituted as well to create point mutations, the most subtle changes a gene can undergo. Genetic changes that have accumulated in a gene over hundreds of millions of years of natural evolution can be mimicked and superseded by several weeks' manipulation in the laboratory.

Genes altered by these techniques can then be reintroduced into the biological systems with which they normally interact. An enzyme having a low affinity for the substrate on which it acts can be engineered to associate avidly with the substrate or even to redirect its attention to novel compounds. A protein that normally is transported to one cellular compartment can be redirected to other sites in the cell. A gene that normally is stimulated to expression by one agent can be made to respond to a completely new signal. In short, by altering the genes that organize a biological system, the molecular biologist can change the usual relations between its elements in ways that show how the system normally works. Many biologists of the future will think of a biological system in terms of a series of well-defined mechanical parts that can be dismantled, engineered and reassembled under the guidance of the molecular mechanic.

It is still far from clear that attempts to reduce complex systems to small and simple components, pushed to an extreme, can provide adequate insights for coming to grips with the great problem biologists confront today: describing the overall functioning of a complex organism. Can the biology of a mammal be understood as simply the sum of a large number of systems, each controlled by a different, well-defined gene? Probably not. A more realistic assessment would be that the interactions of complex networks of genes, gene products and specialized cells underlie many aspects of organismic function. Each gene in an organism has evolved not in isolation but in the context of other genes with which it has interacted continuously over a long period of evolutionary development. Most molecular biologists would concede that they do not yet possess the conceptual tools for understanding entire complex biological systems or processes having multiple interacting components—such as, for example, the process of embryonic development.

Gene cloning has already illuminated another corner of biology: the history of evolutionary development. Human beings, who appeared in their present form only several hundred thousand years ago, are rapidly becoming privy to some of the evolutionary events that, as long as one



CLONING OF A GENE is the central operation of recombinant DNA technology. One approach is diagrammed. In the first step the DNA of a mammalian genome (*1*) is cleaved with a restriction enzyme. Some of the resulting fragments (*2, 3*) may include the gene of interest (*color*). Many copies of a plasmid-cloning vector are cleaved with the same enzyme (*4*). Plasmids and genome fragments are mixed and joined by DNA ligase (*5*), and recombinant plasmids are introduced into bacteria (*6*). The bacteria are plated in a culture dish thinly enough so that each resulting colony (*7*) is a pure clone whose members are descended from a single cell. The cells of a few clones may contain a recombinant plasmid carrying the gene of interest. There is an easy way to find such a clone if this wanted gene's messenger RNA has been identified and can serve as a probe. A sample of the colonies is transferred to a disk of filter paper; the cells are broken open to expose their DNA (*8*). The RNA probe, labeled with a radioactive isotope, is added (*9*). It anneals only to the wanted DNA, and the unannealed probe is removed (*10*). When the filter paper is covered with a photographic emulsion, the radioactive probe makes a spot on the emulsion (*11*), identifying the clone of interest (*12*).

and two billion years ago, began to shape the life-forms that now populate the earth. The steps that generated the first cellular life-forms may never be known, but many of the subsequent changes, memorialized in the DNA of present-day organisms, are being identified by means of the cloning and sequencing of genes. The relatedness of organisms can be established with little ambiguity simply by analysis of their cloned DNA segments. Sophisticated computer programs have been mustered to help analyze these sequences and determine the evolutionary relations involved.

What makes it possible to reconstruct the genealogy of life is the remarkable conservation of certain ancestral sequences over very large evolutionary distances. This conservation has been of great practical advantage to the molecular biologist for another reason. Genes and associated biological problems that are difficult to attack in one organism can be resolved in another organism that is more amenable to manipulation.

There are, for example, certain oncogenes (cancer genes) whose functions are analyzed only with difficulty in human cancer cells. Closely related genes have been identified in the DNA of yeast—an important finding in itself because it indicated that the genes played essential roles in normal cellular physiology long before the appearance of multicellular organisms. The relative simplicity of the yeast cell and the elegant manipulations to which it can be subjected have made experiments possible that are beyond the reach of those working with mammalian cells. The experiments with yeast cells first yielded information on how the oncogenes function; extrapolation of the results to mammalian cells will contribute to the rapidly evolving understanding of the molecular basis of cancer.

The application of cloning techniques to the study of evolution affects understanding of the human species as well. The human species, like others, is composed of a genetically heterogeneous collection of individual organisms. This diversity provides the seeds of future evolution, and the particular versions of genes carried by some people will confer advantage on humankind in future encounters with the forces of natural selection. Versions of certain other genes present in the human gene pool are clearly disadvantageous at present: genes that predispose toward sickle cell disease, atherosclerosis, cancer,

hemophilia and a large number of other metabolic disorders. The genes implicated in these diseases are rapidly being identified and cloned, and with this cloning comes the prospect of making clear diagnoses of genetic predisposition in adults and even in the embryo.

The recognition of genetic diversity and genetic lesions within the human gene pool has stimulated research on ways of repairing defective genes, both in afflicted individuals and in their descendants. Techniques are now available for introducing cloned genes into certain somatic (non-germ-line) tissues such as bone marrow and skin cells. The cloned genes can be intact, healthy versions of genes present only in defective form in the cells of affected individuals. This kind of transfer of genes into somatic cells may reverse, at least partially, the effects of certain genetic lesions. Insertion of cloned genes into the human germ line as well can be contemplated within this decade; the hope is thereby to cure a genetic defect in the descendants of an afflicted individual.

The prospect of such therapeutic intervention has triggered heated debate. There are two broad issues. The first one concerns the human germ line itself. In tampering with it, do human beings cross over an inviolable boundary? Should the human germ line be ringed by a wall to protect its sanctity? And will initial attempts to ameliorate obviously disadvantageous conditions soon be replaced by more ambitious plans to "improve" the human germ line? Genes affecting aspects of intelligence, disposition and body build will surely be identified in the next decade or so, and they could become tempting targets for manipulation.

The other issue transcends the human condition. Is germ-line alteration in general a threat? As described above, genomic alteration is now practical in bacteria, flies, plants and mammals. How will existing ecological interrelations be perturbed by the presence of organisms carrying altered genes in their germ line? The results to date have been reassuring, in that genetically altered organisms have proved to be less viable than their wild-type counterparts and hence unable to affect existing ecosystems substantially. A number of arguments can be mustered, each of which persuades that ecological imbalance is unlikely—and none of which provides total assurance that accidents are impossible.

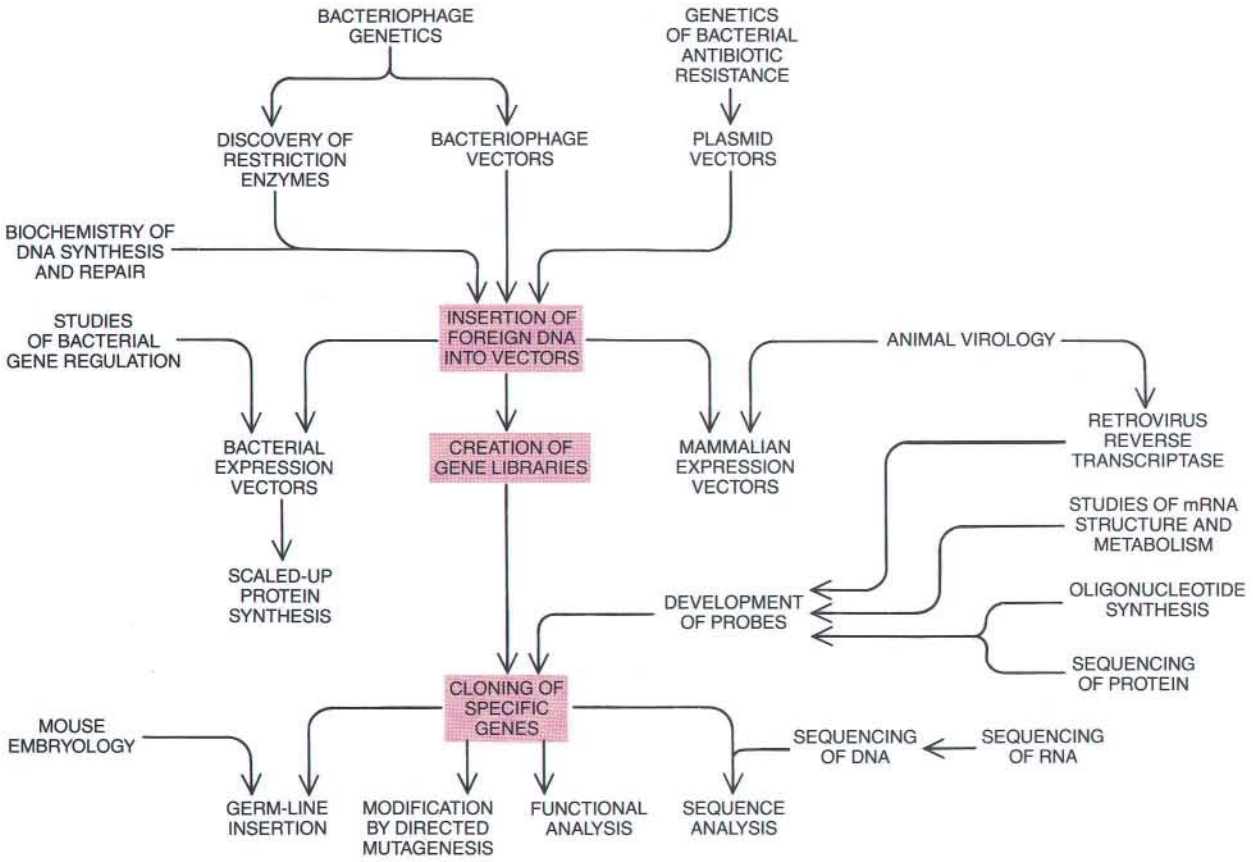
A profound disquiet underlies the consideration of these issues. As physicists did 40 years ago, contemporary biologists have invaded a domain of human innocence. Is life to be redesigned to suit human needs and curiosity? Can—and should—life be described in terms of molecules? For many people, such a description seems to diminish the beauty of nature. For others of us, the beauty and wonder of nature are nowhere more manifest than in the submicroscopic plan of life that has been revealed by the recent discoveries.

FURTHER READING

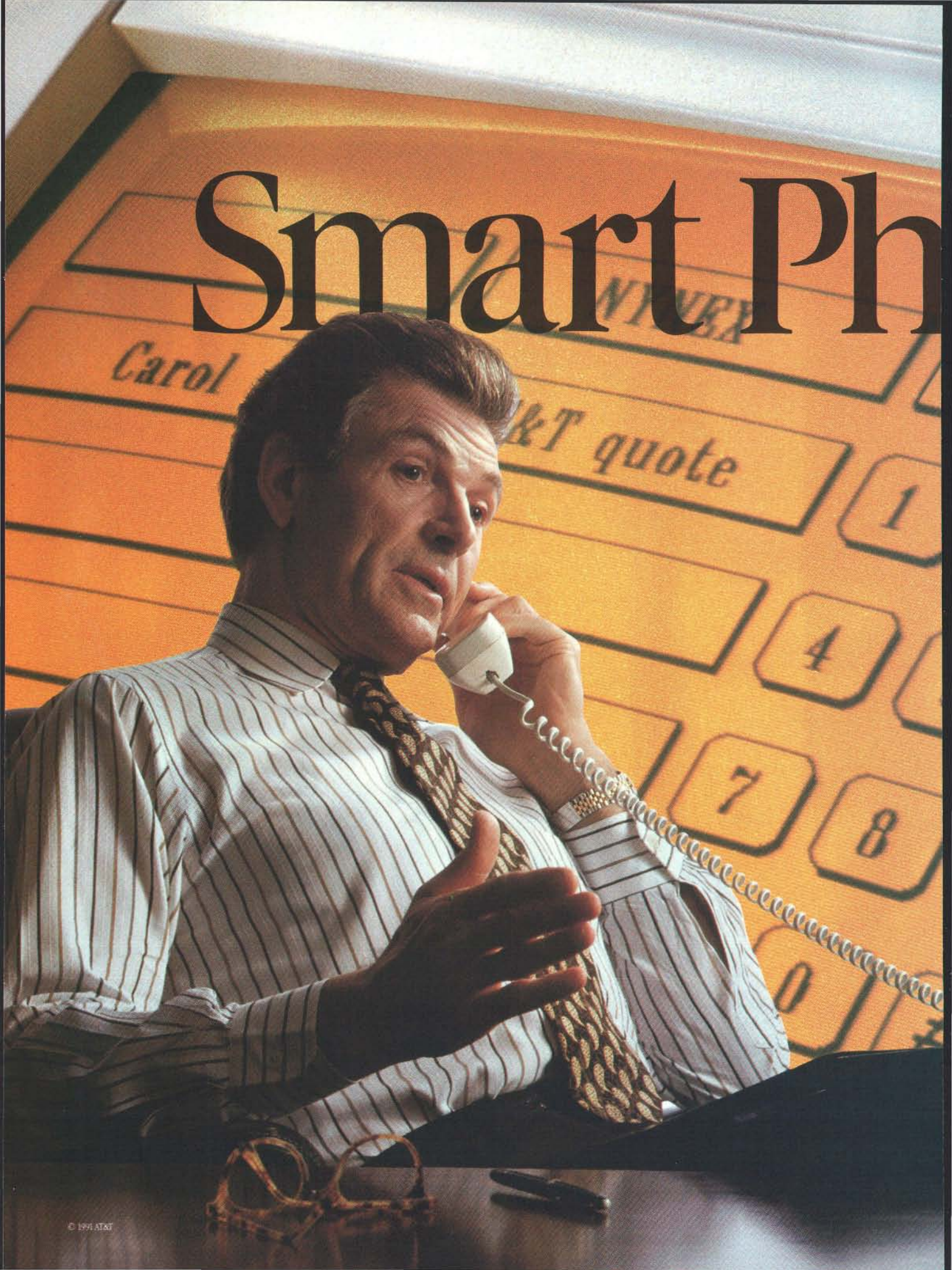
PHAGE AND THE ORIGINS OF MOLECULAR BIOLOGY. Edited by Gunther Stent, J. Cairns, James D. Watson et al. Cold Spring Harbor Laboratory, 1966.

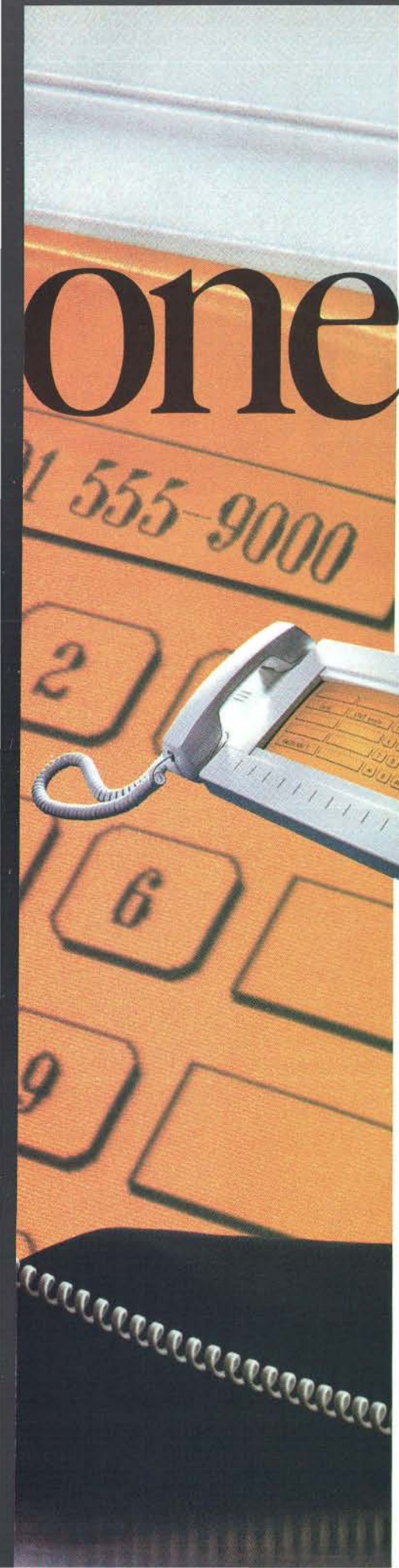
MOLECULAR BIOLOGY OF THE CELL. Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts and James D. Watson. Garland Publishing, Inc., 1983.

RECOMBINANT DNA: A SHORT COURSE. James D. Watson, John Tooze and David T. Kurtz. Scientific American Books, Inc., 1983.



LINE OF DESCENT leading to recombinant DNA technology is traced. It shows how a variety of studies and findings contributed to a succession of new techniques, which led in turn to capabilities that opened up new areas of study. The evolution of technology has been facilitated by the continual development of novel tools and methods.


Smart Ph
Carol
kT quote
1
4
7
8
0
© 1991 AT&T



oneSmarts

Or, How The Dumb Old Phone Will Become A Smart New Tool.

In 1992, AT&T Network Systems will be offering the next step forward in communications: a smart phone. Designed by AT&T Bell Laboratories, our smart phone is the world's first interactive programmable phone. Microchip technology lets you create menus on the touch-sensitive screen to perform customized tasks. For instance, create a button called "Friends." Press and a list of names appears. Press the name you want and the person is automatically dialed. Eventually, the touch-sensitive screen will interact with your bank for transfers and balance inquiries; airline reservation systems to enable you to book flights yourself; and many other databases. Call AT&T Network Systems at **1 800 638-7978, ext. 9210** for a free brochure on how smart phones can give your home or business more smarts.

AT&T Network Systems And Bell Laboratories. Technologies For The Real World.

AT&T
Network Systems

The Unusual Origin of the Polymerase Chain Reaction

A surprisingly simple method for making unlimited copies of DNA fragments was conceived under unlikely circumstances—during a moonlit drive through the mountains of California

by Kary B. Mullis

Sometimes a good idea comes to you when you are not looking for it. Through an improbable combination of coincidences, naiveté and lucky mistakes, such a revelation came to me one Friday night in April 1983 as I gripped the steering wheel of my car and snaked along a moonlit mountain road into northern California's redwood country. That was how I stumbled across a process that could make unlimited numbers of copies of genes, a process now known as the polymerase chain reaction (PCR).

Beginning with a single molecule of the genetic material DNA, the PCR can generate 100 billion similar molecules in an afternoon. The reaction is easy to execute: it requires no more than a test tube, a few simple reagents and a source of heat. The DNA sample that one wishes to copy can be pure, or it can be a minute part of an extremely complex mixture of biological materials. The DNA may come from a hospital tissue specimen, from a single human hair, from a drop of dried blood at the scene of a crime, from the tissues of a mummified brain or from a 40,000-year-old woolly mammoth frozen in a glacier.

In the seven years since that night, applications for the PCR have spread throughout the biological sciences: more than 1,000 reports of its use have been published. Given the impact of the PCR on biological research and its conceptual simplicity, the fact that it lay unrecognized for more than 15 years after all the elements for its implementation were available strikes many observers as uncanny.

The polymerase chain reaction makes life much easier for molecular biologists: it gives them as much of a particular DNA as they want. Casual discussions of DNA molecules sometimes make them sound like easily obtained objects. The truth is that in practice it is difficult to get a well-defined molecule of natural DNA from any organism except extremely simple viruses.

The difficulty resides in the nature of the molecule. DNA is a delicate chain made of four deoxynucleotides: deoxyadenylate (*A*), deoxythymidylate (*T*), deoxyguanylate (*G*) and deoxycytidylate (*C*); the sequence of these bases encodes the genetic information. Rarely does one find a single strand of DNA; usually pairs of strands with complementary sequences form double helixes in which the *A*s in one strand bind with the *T*s in the other, and the *G*s bind with the *C*s [*see illustration on opposite page*]. Inside a cell this DNA helix is surrounded and further coiled by various proteins. When biologists try to isolate a naked DNA chain, the DNA is so long and thin that even mild shearing forces break it at random points along its length. Consequently, if the DNA is removed from 1,000 identical cells, there will be 1,000 copies of any given gene, but each copy will be on a DNA fragment of differing length.

For many years, this problem made it difficult to study genes. Then, in the 1970s, enzymes known as restriction endonucleases were discovered: these enzymes snipped strands of DNA at specific points. The endonucleases made it possible to cut DNA into smaller, sturdier, more identifiable pieces and thereby made it easier to isolate the pieces containing a gene of interest.

By the late 1970s, therefore, molecular biologists were busily studying DNA with endonucleases and with other molecules known as oligonucleotide probes. An oligonucleotide is a short chain of specifically ordered nucleotide bases. Under the right conditions, an oligonucleotide will bind specifically with a complementary sequence of nucleotides in single-strand DNA. Therefore, radioactively labeled, man-made oligonucleotides can serve as probes for determining whether a sample of DNA contains a specific nucleotide sequence or gene. In 1979 the Cetus Corporation in Emeryville, Calif., hired me to synthesize oligonucleotide probes.

By 1983 the charm of synthesizing oligonucleotides for a living had entered a decline—a decline that most of us so employed were happy to witness. The laborious but very quaint chemical art form for making oligonucleotides manually, to which we had grown comfortably numb, had given way to a much less charming but reliable automated technique. It was an immense improvement.

In the aftermath of this minor industrial revolution, we nucleotide chemists

KARY B. MULLIS describes himself as "a generalist with a chemical prejudice." In addition to the polymerase chain reaction, he is also known for having invented a plastic that changes color rapidly when exposed to ultraviolet light. While working as a biochemistry graduate student at the University of California, Berkeley, he published a paper in *Nature* entitled "The Cosmological Significance of Time Reversal." Mullis received his Ph.D. in biochemistry in 1972. After working as a postdoctoral fellow at the University of Kansas Medical School and the University of California, San Francisco, Mullis joined the Cetus Corporation, where he discovered the polymerase chain reaction. In 1986 he became the director of molecular biology at Xytronyx, Inc. Today Mullis works in La Jolla, Calif., as a private consultant on polymerase chain reaction technology and nucleic acid chemistry.

Reprinted from the April 1990 issue

found ourselves successfully underemployed. Laboratory machines, which we loaded and watched, were making almost more oligonucleotides than we had room for in the freezer and certainly more than the molecular biologists—who seemed to be working even more slowly and tediously than we had previously suspected—could use in their experiments. Consequently, in my laboratory at Cetus, there was a fair amount of time available to think and to putter.

I found myself puttering around with oligonucleotides.

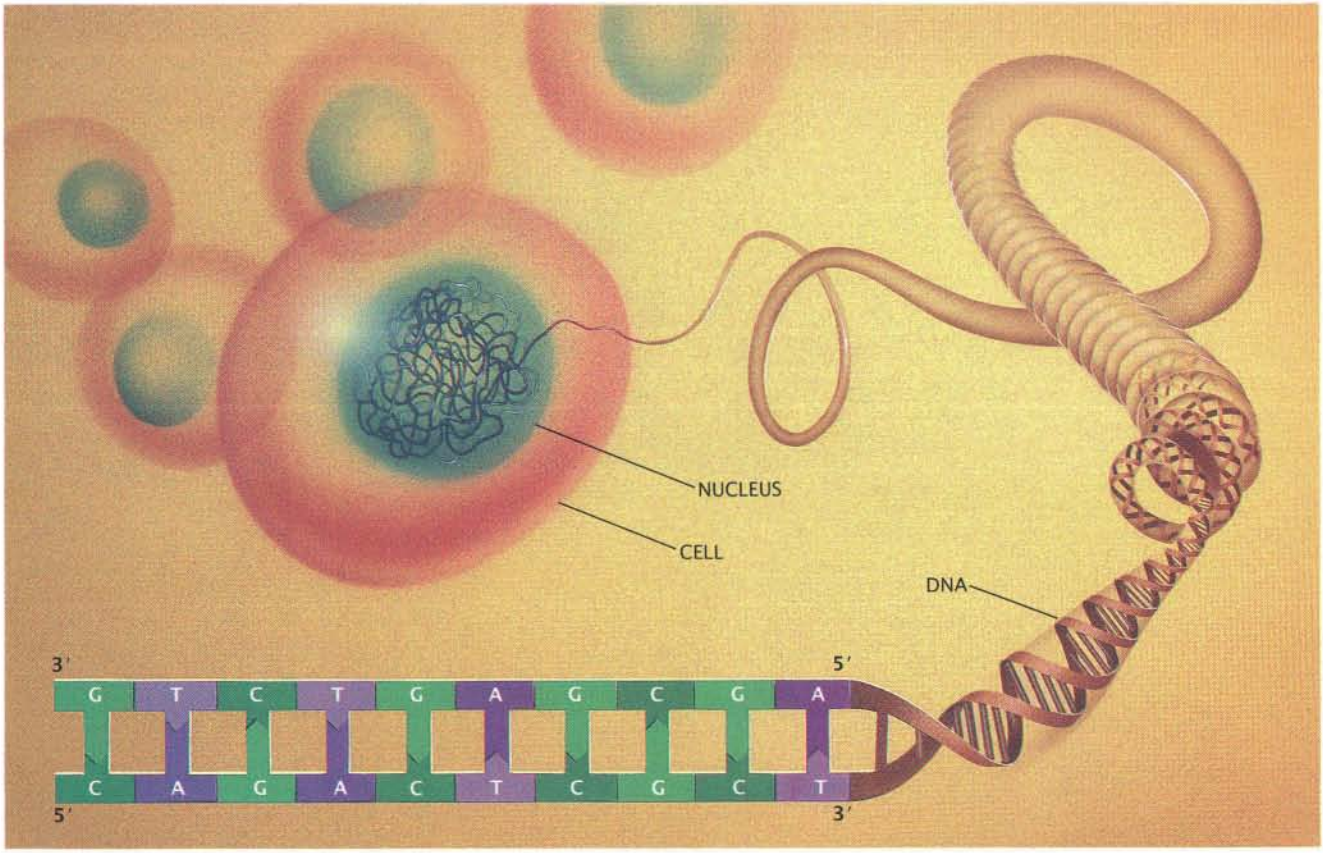
I knew that a technique for easily determining the identity of the nucleotide at a given position in a DNA molecule could be useful, especially if it would work when the complexity of the DNA was high (as it is in human DNA) and when the available quantity of the DNA was small. I did not see why one could not use the enzyme DNA polymerase and a variation of a technique called dideoxy sequencing, and therefore I designed a simple-minded experiment to test the idea.

To understand the approach I had in mind, it is worth reviewing certain facts about DNA. A strand of the molecule has one end that is known, by chemical convention, as three-prime and one end that is five-prime. In a double helix of DNA, the complementary strands are said to be antiparallel, because the three-prime end of one strand pairs with the five-prime of the other strand, and vice versa.

In 1955 Arthur Kornberg of Stanford University and his associates discovered a cellular enzyme called a DNA polymerase. DNA polymerases serve several natural functions, including the repair and replication of DNA. These enzymes can lengthen a short oligonucleotide "primer" by attaching an additional nucleotide to its three-prime end, but only, however, if the primer is hybridized, or bound, to a complementary strand called the template. The surrounding solution must also contain nucleotide triphosphate molecules as building blocks.

The nucleotide that the polymerase attaches will be complementary to the base in the corresponding position on the template strand. For example, if the adjacent template nucleotide is an *A*, the polymerase attaches a *T* base; if the template nucleotide is a *G*, the enzyme attaches a *C*. By repeating this process, the polymerase can extend the primer's three-prime end all the way to the template's five-prime terminus [*see illustration on page 107*]. In a double helix of DNA, each strand serves as a template for the other during replication and repair.

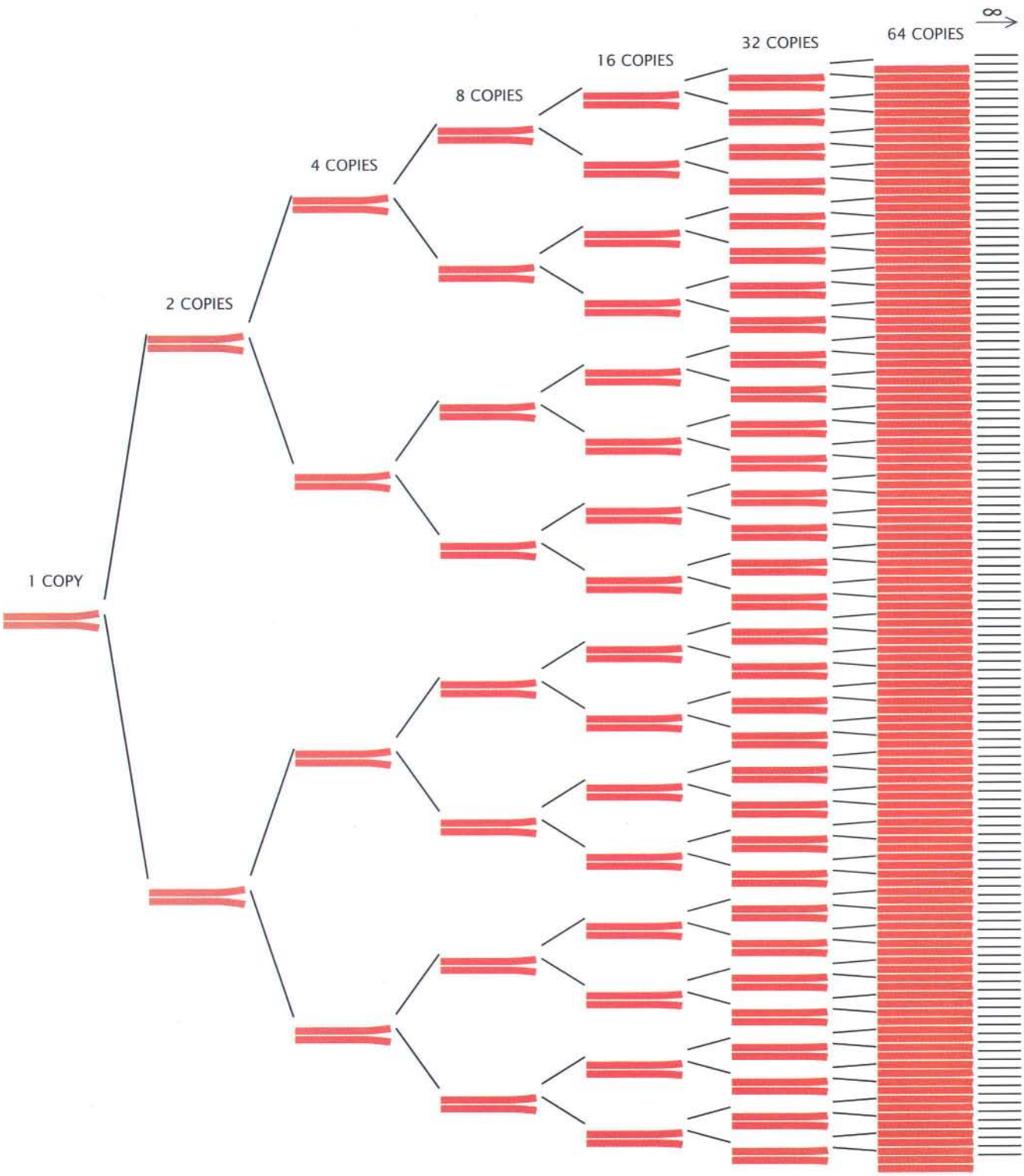
Now for dideoxy sequencing, which is also commonly called the Sanger technique after one of its inventors, Frederick Sanger of the British Medical Research Council Laboratory of Molecular Biology. This technique uses a DNA polymerase, template strands, primers, nucleotide triphosphates and special dideoxynucleotide triphosphates (ddNTPs) to determine DNA sequences. Like ordinary nucleotides, ddNTPs can be attached to growing primers by polymerases; however, a ddNTP will "cap" the three-prime end of a primer and prevent the addition of any more bases. The Sanger technique produces primers that have been lengthened to varying extents and then capped by a ddNTP. By arranging these fragments according to length and by knowing which ddNTPs have been added, an investigator can determine the sequence of bases in the template strand. For example, if a dideoxyadenine (ddA) base were added at a given position, the corresponding complementary base in the template would be



DNA consists of two strands of linked nucleotides: deoxyadenylates (*A*s), deoxythymidylates (*T*s), deoxyguanylates (*G*s) and deoxycytidylates (*C*s). The sequence of nucleotides in one strand is complementary to the sequence in the other: *A*s opposite *T*s and *G*s opposite *C*s. This complementarity binds the strands together. Each strand has a three-prime and a five-prime end. Because their orientations oppose one another, the strands are said to be antiparallel.

a *T*; the addition of a dideoxyguanine (ddG) implies the presence of a *C* in the template. In the modified version of this technique that I was contemplating, I would use only polymerases, templates, ddNTPs and primer molecules—that is, I would omit the ordinary nucleotide triphosphates from the mixture. Extension of the primers would therefore terminate immediately after the addition of one base from a ddNTP to the chain. If I knew which ddNTP had been added to the primers, I would also know the identity of the corresponding base in the template strand. In this way, I could deduce the identity of a base in the template strand adjacent to the site where the primer binds.

What I did not realize at the time was that there were many good reasons



POLYMERASE CHAIN REACTION is a simple technique for copying a piece of DNA in the laboratory with readily available reagents. Because the number of copies increases exponentially, more than 100 billion can be made in a few hours.

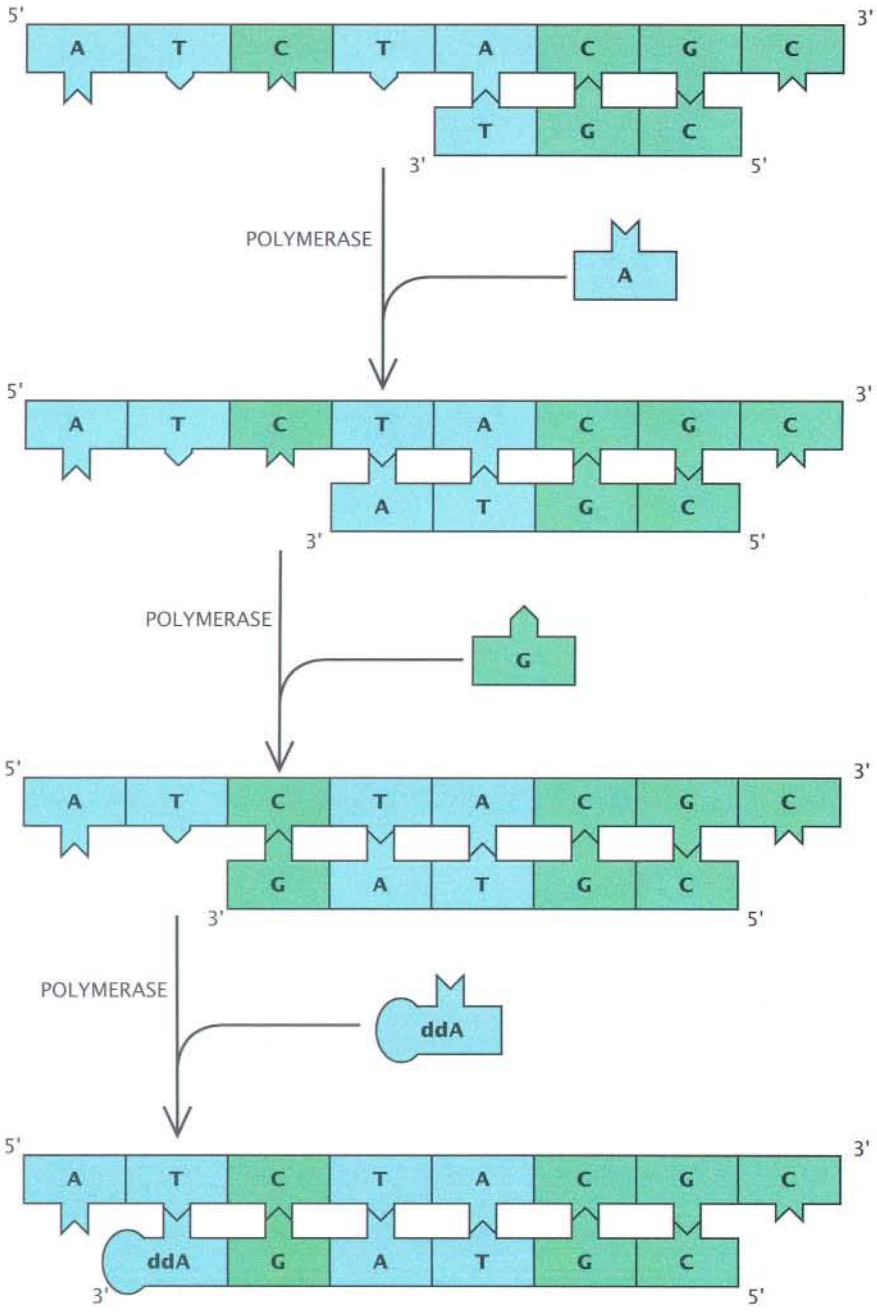
why my sequencing idea could not work. The problem was that oligonucleotides sometimes hybridize with DNA sequences other than those intended; these unavoidable pairings would have made my results ambiguous. Even in the hands of those skilled in the art of careful hybridization, it was impossible to bind oligonucleotides to whole human DNA with sufficient specificity to get anything even approaching a meaningful result.

It was because of this limitation that researchers had resorted to more difficult procedures for looking at human DNA. For instance, restriction enzymes could be employed to cleave the DNA sample into various fragments that could be separated from one another by electrophoresis; in this way, the sample could be "purified," to some extent, of all DNA except the target fragment before the hybridization of oligonucleotide probes. This approach reduced erroneous hybridizations sufficiently to provide meaningful data, but just barely. Moreover, this procedure was lengthy and would not work on degraded or denatured samples of DNA.

Another technique that was much too lengthy for routine DNA analysis involved cloning. A human DNA sequence of interest could be cloned, or copied, into a small ring of DNA called a plasmid. Copies of this plasmid and the targeted sequence could then be produced in bacteria, and sequence information could be obtained by oligonucleotide hybridization and dideoxy sequencing. In the early 1980s dideoxy sequencing of cloned DNA was the method by which most human DNA sequence information had been obtained.

In proposing my simple-minded experiment, I was implicitly assuming that no such cloning or other step would be necessary to detect specific human DNA sequences by a single oligonucleotide hybridization. In token defense of my misguided puttering, I can point out that a group down the hall led by Henry A. Erlich, one of Cetus's senior scientists, was trying another method based on the hybridization of a single oligonucleotide to a human DNA target. No one laughed out loud at Henry, and we were all being paid regularly. In fact, we were being paid enough to lead some of us to assume, perhaps brashly, that we were somewhere near the cutting edge of DNA technology.

One Friday evening late in the spring, I was driving to Mendocino County with a chemist friend. She was asleep.



DNA POLYMERASE can lengthen a short strand of DNA, called an oligonucleotide primer, if the strand is bound to a longer "template" strand of DNA. The polymerase does this by adding the appropriate complementary nucleotide to the three-prime end of the bound primer. If a dideoxynucleotide triphosphate (ddNTP) such as dideoxyadenine (ddA) is added, however, no further extension is possible, because the three-prime end of the ddA will not link to other nucleotides.

U.S. 101 was undemanding. I liked night driving; every weekend I went north to my cabin and on the way sat still for three hours in the car, my hands occupied, my mind free. On that night I was thinking about my proposed DNA-sequencing experiment.

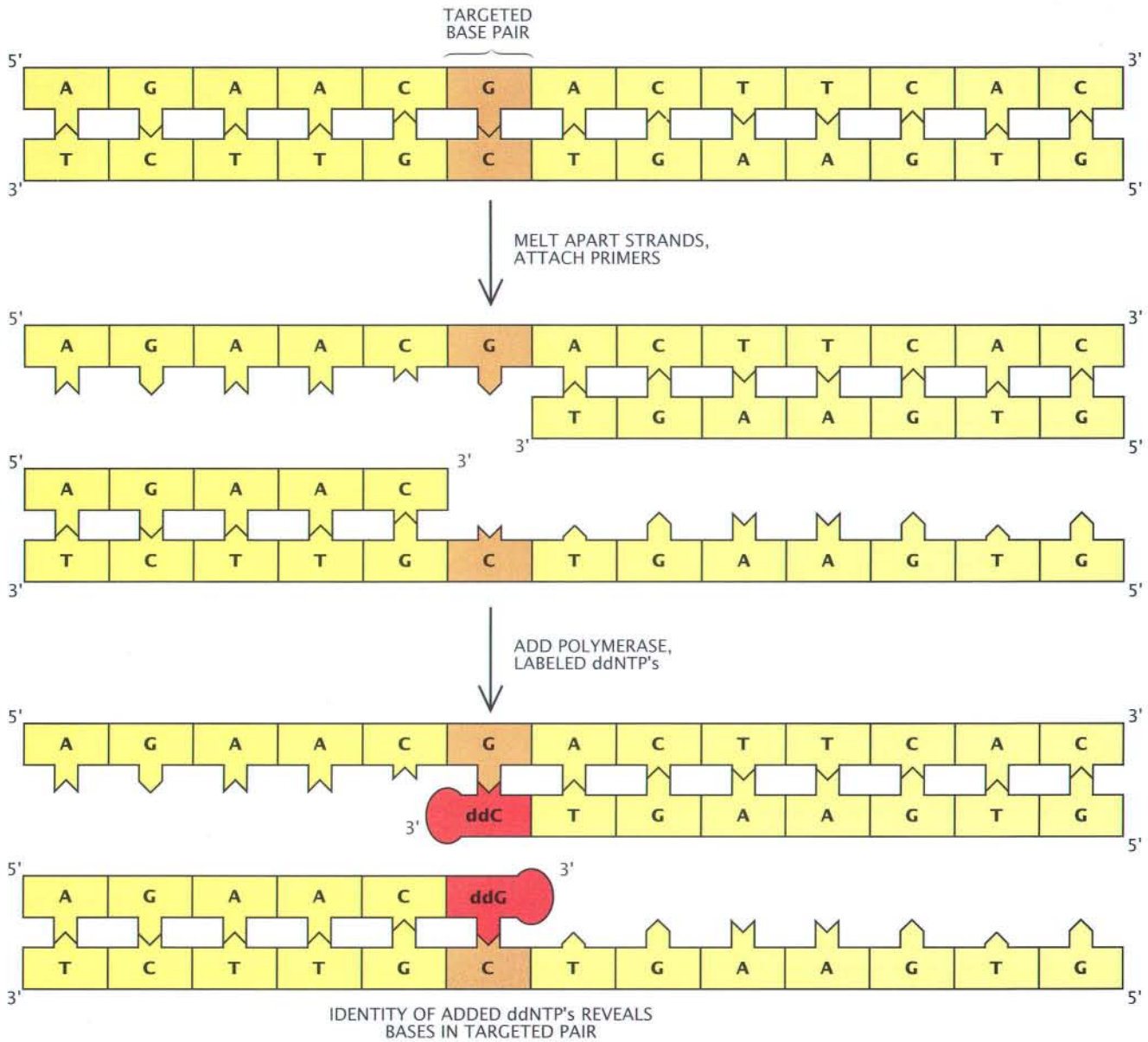
My plans were straightforward. First I would separate a DNA target into single strands by heating it. Then I would hybridize an oligonucleotide to a complementary sequence on one of the strands. I would place portions of this DNA mixture into four different tubes. Each tube would contain all four types of ddNTPs, but in each tube a different type of ddNTP would be radioactively labeled. Next I would add DNA polymerase, which would extend the hybridized oligonucleotides in each tube

by a single ddNTP. By electrophoresis, I could separate the extended oligonucleotides from the residual ddNTPs; by identifying which radioactively labeled ddNTP had been incorporated into the oligonucleotide, I could determine the corresponding complementary base in the target strand. Simple.

Near Cloverdale, where California 128 branches northwest from U.S. 101 and winds upward through the coastal range, I decided the determination would be more definitive if, instead of just one oligonucleotide, I used two. The two primers would bracket the targeted base pair I hoped to identify. By making the oligonucleotides of different sizes, I would be able to distinguish them from each other. By directing one oligonucleotide to each strand of the sample DNA target, I could get complementary sequencing information about both strands. Thus the experiment would contain an internal control at no extra inconvenience [*see illustration below*].

Although I did not realize it at that moment, with the two oligonucleotides poised in my mind, their three-prime ends pointing at each other on opposite strands of the gene target, I was on the edge of discovering the polymerase chain reaction. Yet what I most felt on the edge of was the mountain road.

That night the air was saturated with moisture and with the scent of flowering buckeye. The reckless white stalks poked from the roadside into the glare of my headlights. I was thinking about the new ponds I was digging on my property, while also hypothesizing about things that might go wrong



TO DETERMINE THE IDENTITY of a targeted base pair in a piece of DNA, the author hoped to apply a variation on a technique called dideoxy sequencing. First, two primers would be bound to the opposing strands in the DNA at sites flanking the targeted pair. DNA polymerase and dideoxynucleotide triphosphates (ddNTPs) would then be added to the mixture, allowing each of the primers to be extended by only one base. The identity of the added ddNTP bases would reveal what the complementary targeted bases were. The technique could work with only one primer, but the use of two would provide a control for checking the results. Planning this experiment led the author to the polymerase chain reaction.

with my base-sequencing experiment.

From my postdoctoral days in Wolfgang Sadee's laboratory at the University of California at San Francisco, where John Maybaum was devising clinical assays for nucleotides, I remembered that my DNA samples might contain stray traces of nucleotide triphosphates. It would complicate the interpretation of the gel, I figured, if stray nucleotides introduced with the sample added themselves to the three-prime end of the primers before the planned addition of the labeled ddNTPs.

One thought was to destroy any loose nucleotide triphosphates in the sample with alkaline phosphatase, a bacterial enzyme. This enzyme would chew the reactive phosphate groups off any nucleotide triphosphates, thereby rendering them inert to a polymerase reaction. Yet I would then somehow have to eliminate the phosphatase from the sample, or else it would also destroy the ddNTPs when I added them. Normally one can deactivate unwanted enzymes by heating them and altering their essential shape; I believed, however, bacterial alkaline phosphatase could refold itself into its original form. I therefore rejected alkaline phosphatase as an answer to the problem.

I was, in fact, mistaken. Much later I learned that alkaline phosphatase can be irreversibly denatured by heating if no zinc is present in the solution. As it turned out, my mistake was extraordinarily fortunate: if I had known better, I would have stopped searching for alternatives.

Every mile or so another potential solution arose but fell short. Then, as I began the descent into Anderson Valley, I hit on an idea that appealed to my sense of aesthetics and economy: I would apply the same enzyme, DNA polymerase, twice—first to eliminate the extraneous nucleotide triphosphates from the sample, then to incorporate the labeled ddNTPs.

I reasoned that if there were enough nucleotides in the sample to interfere with the experiment, there would also be enough for the DNA polymerase to act on. By running the sample through a kind of preliminary mock reaction with oligonucleotide primers and polymerase but without ddNTPs, I could easily deplete any nucleotides in the mixture by incorporating them into the extending oligonucleotides. Then, by raising the temperature of the sample, I could separate the extended oligonucleotides from the DNA targets. True, the extended oligonucleotides would still be in the sample, but because there would be far more unextended primers than extended ones in the mixture, the DNA targets would probably hybridize with unextended primers when the mixture cooled. I could then add ddNTPs and more polymerase to perform my sequencing experiment.

Yet some questions still nagged at me. Would the oligonucleotides extended by the mock reaction interfere with the subsequent reactions? What if they had been extended by many bases, instead of just one or two? What if they had been extended enough to create a sequence that included a binding site for the other primer molecule? Surely that would cause trouble...

No, far from it! I was suddenly jolted by a realization: the strands of DNA in the target and the extended oligonucleotides would have the same base sequences. In effect, the mock reaction would have doubled the number of DNA targets in the sample!

Suddenly, for me, the fragrance of the flowering buckeye dropped off exponentially.

Under other circumstances, I might not have recognized the importance of this duplication so quickly. Indeed, the idea of repeating a procedure over and over again might have seemed unacceptably dreary. I had been spending a lot of time writing computer programs, however, and had become familiar with reiterative loops—procedures in which a mathematical operation is repeatedly applied to the products of earlier iterations. That experience had taught me how powerful reiterative exponential growth processes are. The DNA replication procedure I had imagined would be just such a process.

Excited, I started running powers of two in my head: two, four, eight, 16, 32...I remembered vaguely that two to the tenth power was about 1,000 and that therefore two to the twentieth was around a million. I stopped the car at a turnout overlooking Anderson Valley. From the glove compartment I pulled a pencil and paper—I needed to check my calculations. Jennifer, my sleepy passenger, objected groggily to the delay and the light, but I exclaimed that I had discovered something fantastic. Unimpressed, she went back to sleep. I confirmed that two to the twentieth power really was over a million and drove on.

About a mile farther down the road I realized something else about the products of the reaction. After a few rounds of extending the primers, dissociating the extension products, rehybridizing new primers and extending them, the length of the exponentially accumulating DNA strands would be fixed because their ends would be sharply defined by the five-prime ends of the oligonucleotide primers. I could replicate larger fragments of the original DNA sample by designing primers that hybridized farther apart on it. The fragments would always be discrete entities of a specified length.

I stopped the car again and started drawing lines of DNA molecules hybridizing and extending, the products of one cycle becoming the templates for the next in a chain reaction. Jennifer protested again from the edge of sleep. "You're not going to believe this," I crowed. "It's incredible."

She refused to wake up. I proceeded to the cabin without further stops. The deep end of Anderson Valley is where the redwoods start and where the "ne'er-do-wells" have always lived. My discovery made me feel as though I was about to break out of that old valley tradition. It was difficult for me to sleep that night with deoxyribonuclear bombs exploding in my brain.

Yet in the morning I was too tired not to believe that someone, somewhere, must have tried this idea already. Thousands of investigators had, for various reasons, extended single oligonucleotides with polymerases; surely someone would have noticed the possibility of a polymerase chain reaction. But if it had worked, I was sure I would have heard about it: people would have been using it all the time to amplify, or multiply, DNA fragments.

Back at Cetus on Monday I asked one of the librarians, George McGregor, to run a literature search on DNA polymerase. Nothing relevant to amplification turned up. For the next few weeks, I described the idea to anyone who would listen. No one had heard of its ever being tried; no one saw any good reason why it would not work; and yet no one was particularly enthusiastic about it. In the past, people had generally thought my ideas about DNA were off the wall, and sometimes after a few days I had agreed with them. But this time I knew I was on to something.

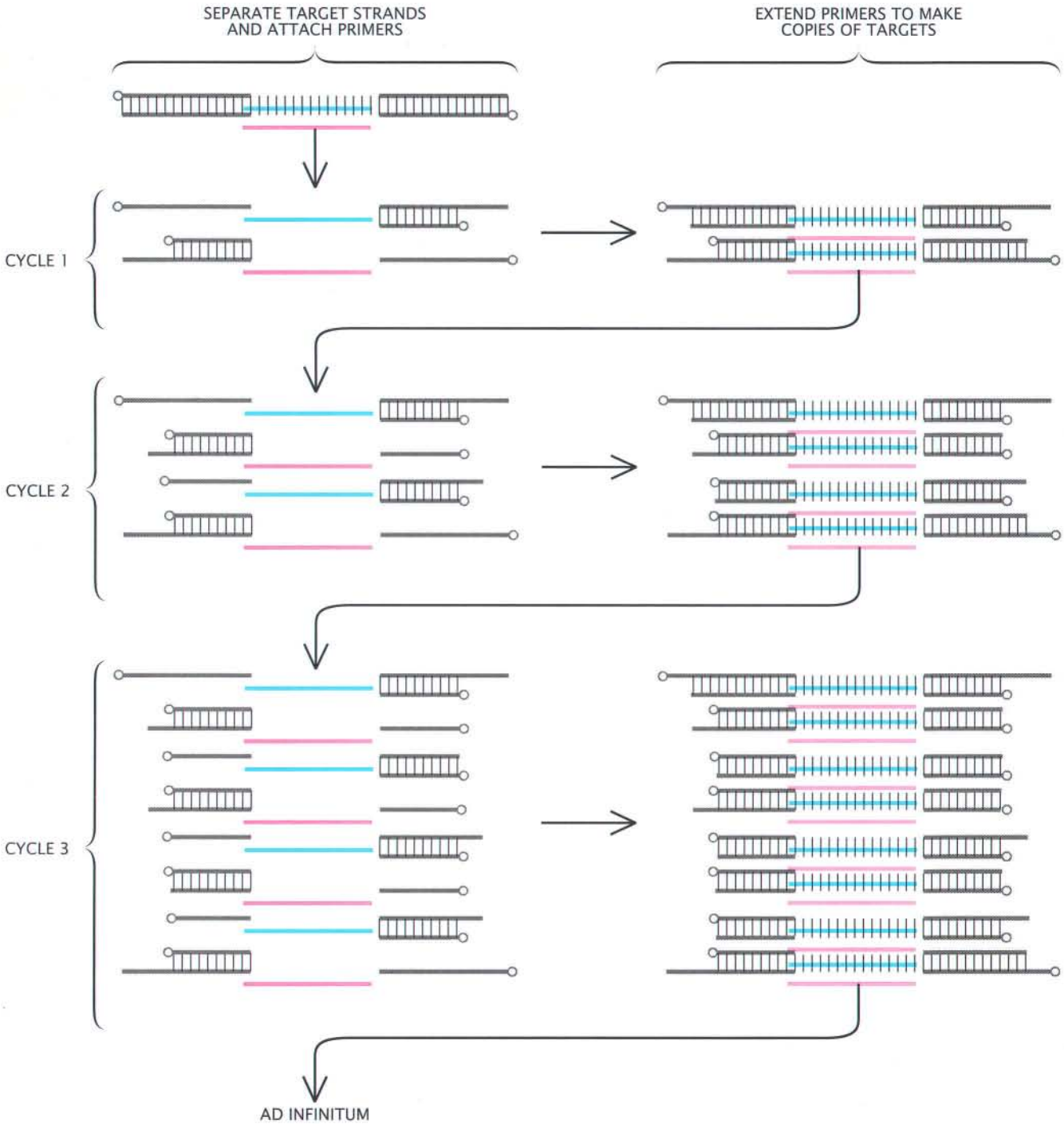
Years ago, before biotechnology—when being a genetic engineer meant that you, your dad and his dad all drove trains—our building at Cetus had been owned by the Shell Development Company. Our laboratory space, whose rear windows looked grandly out on the Berkeley hills, had given birth to the "No-Pest Strip." It did not escape my notice that the PCR might

someday travel as far as its sibling invention, that distinctively scented piece of yellow plastic.

Months passed as I prepared for my first experiment to verify whether the PCR would work. I had to make many educated guesses about what buffer solutions to use, what the relative and absolute concentrations of the reactants should be, how much to heat and cool the mixtures, how long the mixtures should run and so on. Some of Kornberg's early papers on DNA polymerase helped. To run the experiment, I selected a 25-base-pair target fragment of a plasmid and two oligonucleotide primers that were 11 and 13 bases long, respectively.

When everything was ready, I ran my favorite kind of experiment: one involving a single test tube and producing a yes or no answer. Would the PCR amplify the DNA sequence I had selected? The answer was yes.

Walking out of the lab fairly late in the evening, I noticed that Albert Halluin, the patent attorney for Cetus, was still in his office. I told him that I had invented something and described the PCR. Al was the first person, out of



POLYMERASE CHAIN REACTION is a cyclic process; with each cycle, the number of DNA targets doubles. The strands in each targeted DNA duplex are separated by heating and then cooled to allow primers to bind to them. Next DNA polymerases extend the primers by adding nucleotides to them. Duplicates of the original DNA-strand targets are thus produced.

maybe a hundred to whom I had explained it, who agreed that it was significant. He wanted to see the autoradiogram showing the experimental data right away; it was still wet.

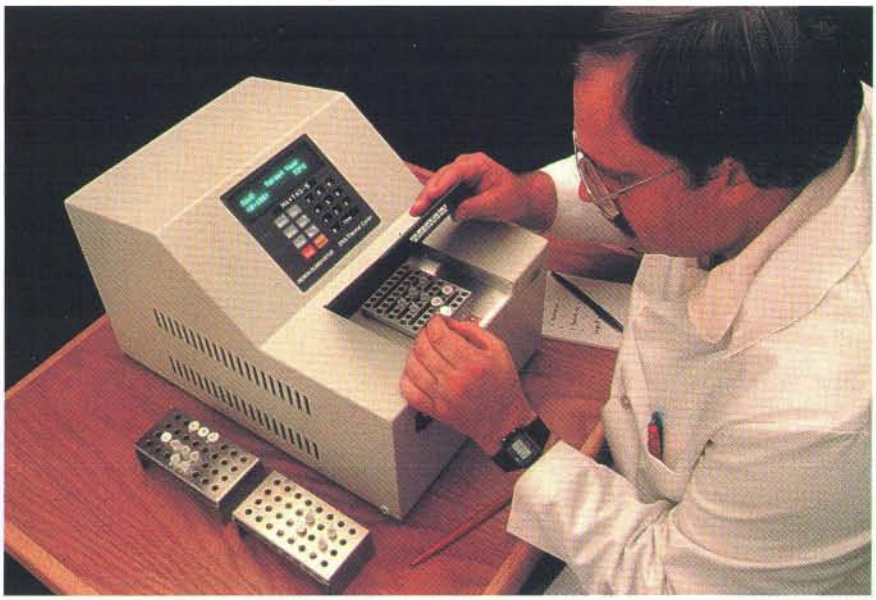
Some people are not impressed by one-tube experiments, but Al was not noticeably skeptical. Patent attorneys, after all, have a vested interest in inventions. He had followed my explanation of the process in his office and agreed that it made sense. Now in the lab he was even a little excited and suggested that I get to work on the experiment and write a patent disclosure. As he left, he congratulated me.

For the next few months, I continued to study and refine the PCR with the help of Fred A. Faloona, a young mathematics wizard whom I had met through my daughter. Fred had helped me with the first PCR experiment by cycling the DNA mixture—in fact, that had been his very first biochemistry experiment, and he and I celebrated on the night of its success with a few beers.

In the following months we confirmed that the PCR would work on larger and larger fragments of plasmid DNA. Eventually we obtained some human DNA from Henry Erlich's laboratory and produced evidence for the amplification of a fragment from a single-copy gene.

Today many of the initial hitches or inefficiencies of the PCR have been worked out. Several slightly different protocols are now in use. I usually recommend that the DNA samples be cycled between temperatures of about 98 degrees Celsius, just below boiling, and about 60 degrees C. These cycles can be as short as one or two minutes; during each cycle the number of DNA target molecules doubles. The primers are usually from 20 to 30 bases long. One of the most important improvements in the process is the use of a particular DNA polymerase originally extracted from the bacterium *Thermus aquaticus,* which lives in hot springs. The polymerase we had originally used was easily destroyed by heat, and so more had to be added during each cycle of the reaction. The DNA polymerase of *Thermus aquaticus,* however, is stable and active at high temperatures, which means that it only needs to be added at the beginning of the reaction. This high-temperature polymerase is now produced conveniently by genetically engineered bacteria.

The virtually unlimited amplification of DNA by the PCR was too unprecedented to be accepted readily. No one was prepared for a process that provided all the DNA one could want. The reaction seemed self-evident to Fred and to me because it was our toy. For most people, it took some getting used to.



MACHINE that performs the polymerase chain reaction is shown as samples of DNA are loaded. Such devices are rapidly becoming common fixtures in laboratories.

In the spring of 1984, while working on the patent, I presented a poster describing the PCR at the annual Cetus Scientific Meeting. These meetings were always fun, because Cetus had some first-rate scientific advisers, and I was looking forward to talking with them about my invention.

Yet nobody seemed to be interested in my poster, and I felt increasingly anxious. People would glance at it and keep walking. Finally, I noticed Joshua Lederberg, president of the Rockefeller University, nearby, and I snared him into looking at my results. Josh looked the poster over carefully and then turned his enormous head, the Nobel-laureated head, the head that had deduced in 1946 that bacteria could have sexual intercourse. "Does it work?" He seemed amused.

Pleased, I confirmed that it did, and we talked for a long time. At one point he mentioned that about 20 years previously, after Kornberg had discovered DNA polymerase, the two of them had considered the notion that the enzyme could somehow be harnessed to make large quantities of DNA. They had not figured out exactly how to do it, however. I reminded him that oligonucleotides were not readily available at that time and that there was hardly any DNA sequence information either.

But he looked back at my poster with an expression that I have almost come to expect. I think that Josh, after seeing the utter simplicity of the PCR, was perhaps the first person to feel what is now an almost universal first response to it among molecular biologists and other DNA workers: "Why didn't I think of that?" And nobody really knows why; surely I don't. I just ran into it one night.

FURTHER READING

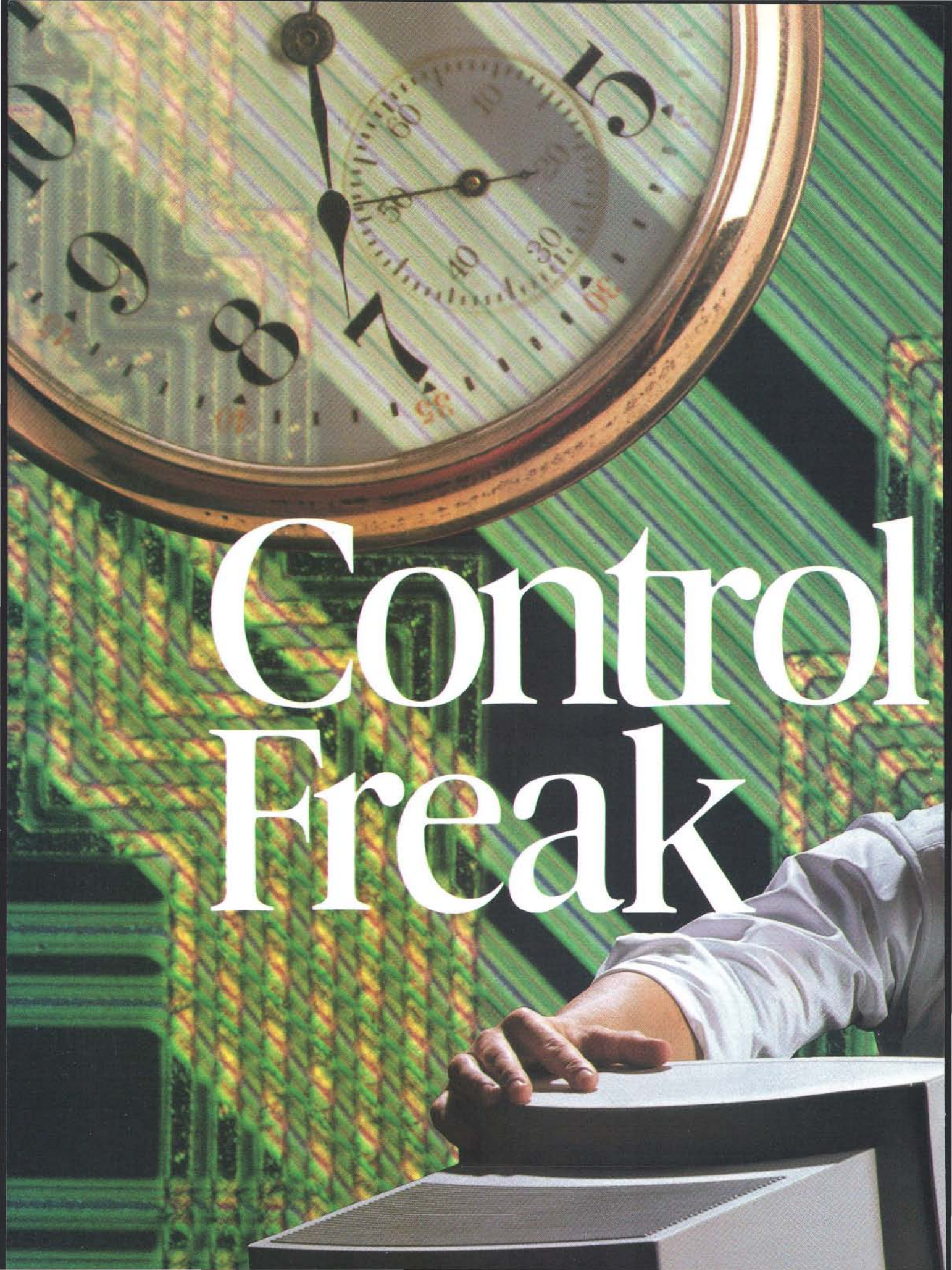
SPECIFIC ENZYMATIC AMPLIFICATION OF DNA IN VITRO: THE POLYMERASE CHAIN REACTION. Kary Mullis, Fred Faloona, Stephen Scharf, Randall Saiki, Glenn Horn and Henry Erlich in *Cold Spring Harbor Symposia on Quantitative Biology,* Vol. 51, No. 1, pages 263–273; 1986.

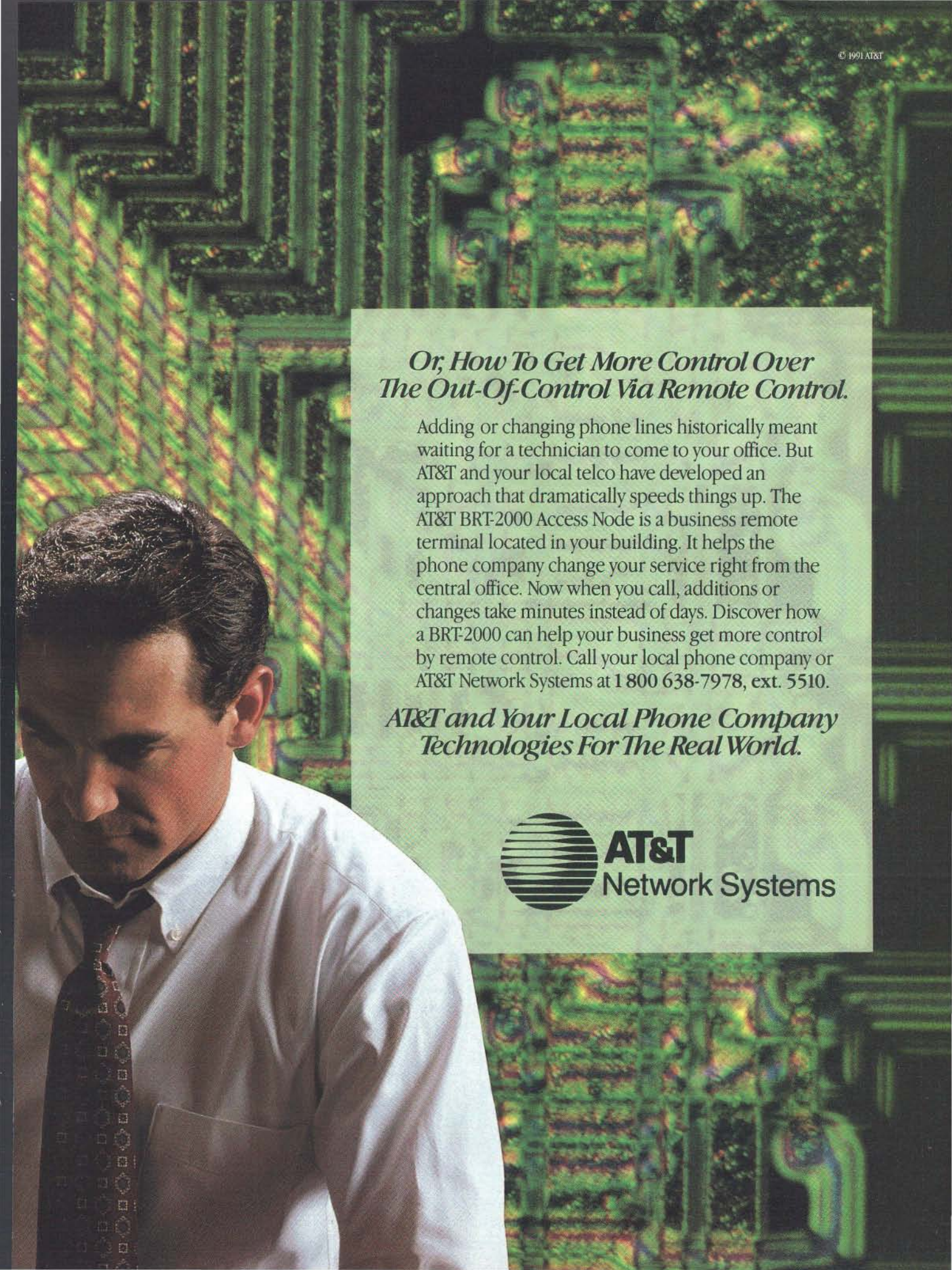
SPECIFIC SYNTHESIS OF DNA IN VITRO VIA A POLYMERASE-CATALYZED CHAIN REACTION. Kary B. Mullis and Fred A. Faloona in *Methods in Enzymology,* Vol. 155, Part F, pages 335–350; 1987.

AMPLIFICATION OF HUMAN MINISATELLITES BY THE POLYMERASE CHAIN REACTION: TOWARDS DNA FINGERPRINTING OF SINGLE CELLS. Alec J. Jeffreys, Victoria Wilson, Rita Neumann and John Keyte in *Nucleic Acids Research,* Vol. 16, pages 10953–10971; 1988.

DNA SEQUENCING WITH THERMUS AQUATICUS DNA POLYMERASE AND DIRECT SEQUENCING OF POLYMERASE CHAIN REACTION-AMPLIFIED DNA. M. A. Innis, K. B. Myambo, D. H. Gelfand and Mary Ann D. Brow in *Proceedings of the National Academy of Sciences,* Vol. 85, No. 24, pages 9436–9440; December 1988.

THE POLYMERASE CHAIN REACTION. T. J. White, Norman Arnheim and H. A. Erlich in *Trends in Genetics,* Vol. 5, No. 6, pages 185–188; June 1989.


Control
Freak


© 1991 AT&T
Or, How To Get More Control Over The Out-Of-Control Via Remote Control.
Adding or changing phone lines historically meant waiting for a technician to come to your office. But AT&T and your local telco have developed an approach that dramatically speeds things up. The AT&T BRT-2000 Access Node is a business remote terminal located in your building. It helps the phone company change your service right from the central office. Now when you call, additions or changes take minutes instead of days. Discover how a BRT-2000 can help your business get more control by remote control. Call your local phone company or AT&T Network Systems at 1 800 638-7978, ext. 5510.
AT&T and Your Local Phone Company Technologies For The Real World.
AT&T
Network Systems

Continental Drift

In 1912 Alfred Wegener proposed that the continents had originated in the breakup of one supercontinent. His idea has not been widely accepted, but new evidence suggests that the principle is correct

by J. Tuzo Wilson

Geology has reconstructed with great success the events that lie behind the present appearance of much of the earth's landscape. It has explained many of the observed features, such as folded mountains, fractures in the crust and marine deposits high on the surface of continents. Unfortunately, when it comes to fundamental processes—those that formed the continents and ocean basins, that set the major periods of mountain-building in motion, that began and ended the ice ages—geology has been less successful. On these questions there is no agreement, in spite of much speculation. The range of opinion divides most sharply between the position that the earth has been rigid throughout its history, with fixed ocean basins and continents, and the idea that the earth is slightly plastic, with the continents slowly drifting over its surface, fracturing and reuniting and perhaps growing in the process. Whereas the first of these ideas has been more widely accepted, interest in continental drift is currently on the rise. In this article I shall explore the reasons for the change.

The subject is large and full of pitfalls. The reader should be warned that I am not presenting an accepted or even a complete theory but one man's view of fragments of a subject to which many are contributing and about which ideas are rapidly changing and developing. If it is conceded that much of this discourse is speculation, then it should also be added that many of the accepted ideas have in fact been speculations also.

In the past, several different theories of continental drift have been advanced, and each has been shown to be wrong in some respects. Until it is indisputably established that such movements in the earth's crust are impossible, however, a multitude of theories of continental drift remain to be considered. Although there is only one pattern for fixed continents and a rigid earth, many patterns of continental migration are conceivable.

The traditional rigid-earth theory holds that the earth, once hot, is now cooling, that it became rigid at an early date and that the contraction attendant on the cooling process creates compressive forces that, at intervals, squeeze up mountains along the weak margins of continents or in deep basins filled with soft sediments. This view, first suggested by Isaac Newton, was quantitatively established during the 19th century to suit ideas then prevailing. It was found that an initially hot, molten earth would cool to its present temperature in about 100 million years and that, in so doing, its circumference would contract by at least tens and perhaps hundreds of miles. The irregular shape and distribution of continents presented a puzzle, but, setting this aside, it was thought that the granitic blocks of the continents had differentiated from the rest of the crustal rock and had frozen in place at the close of the first, fluid chapter of the earth's history. Since then the continents had been modified in situ, without migrating.

This hypothesis, in its essentials, still has many adherents. They include most geologists, with notable exceptions among those who work around the margins of the southern continents. The validity of the underlying physical theory is defended by some physicists. On the other hand, a number of formidable objections have been raised by those who have studied radioactivity, ancient climates, terrestrial magnetism and, most recently, submarine geology. Many biologists have seen another problem. They argue that although the evolution and migration of later forms of life—particularly since the advent of mammals—could be satisfactorily traced on the existing pattern of continents, the distribution of earlier forms required either land bridges across the oceans—the origin and disappearance of which are difficult to explain—or a different arrangement of the continents.

The discovery of radioactivity altered the original concept of the contraction theory without absolutely invalidating it. In the first place, the age of the earth could be reliably determined from knowledge of the rate at which the unstable isotopes of various elements decay and by measurement of the ratios of daughter to parent isotopes present in the rocks. These studies showed the earth to be much older than had been imagined, perhaps 4.5 billion years old. Dating of the rocks indicated that the continents are zoned and have apparently grown by accretion over the ages. Finally, it was found that the decay of

J. TUZO WILSON was professor of geophysics and director of the Institute of Earth Sciences at the University of Toronto when he wrote this article in 1963. Wilson first studied at Toronto, taking an A.B. there in 1930, and then at the University of Cambridge, where he received an A.B. and M.A. in 1932. In 1936 he acquired a Ph.D. from Princeton University and for the next three years was an assistant geologist with the Geological Survey of Canada. His service with the Royal Canadian Engineers for the duration of World War II brought him an Order of the British Empire and the Legion of Merit. He became professor of geophysics at Toronto in 1946 and director of the Institute in 1960. From 1957 to 1960 Wilson was president of the International Union of Geodesy and Geophysics. He served as president of the Ontario Science Center from 1974 to 1985.

Reprinted from the April 1963 issue

ROBESON CHANNEL separating northwestern Greenland (*upper right*) from Ellesmere Island (*foreground*) marks the Wegener fault (named by the author for the German meteorologist who predicted its existence and a great lateral displacement along the length of the channel). Not yet fully mapped, it probably joins a known fault farther southwest.

AGE OF ATLANTIC ISLANDS, as indicated by the age of the oldest rocks found on them, apparently tends to increase with increasing distance from the Mid-Atlantic Ridge. The numbers associated with the islands give these ages in millions of years. Geologists divide Iceland into three areas of different ages, the central one being the youngest. The Rio Grande and Walvis ridges are lateral ridges that may have formed as a result of the drifting apart of Africa and South America. Other lateral ridges are shown. Islands that have active volcanoes are represented by black triangles; most of them lie on or near the Mid-Atlantic Ridge. The extension of the ridge into Baffin Bay is postulated. Broken colored lines are faults.

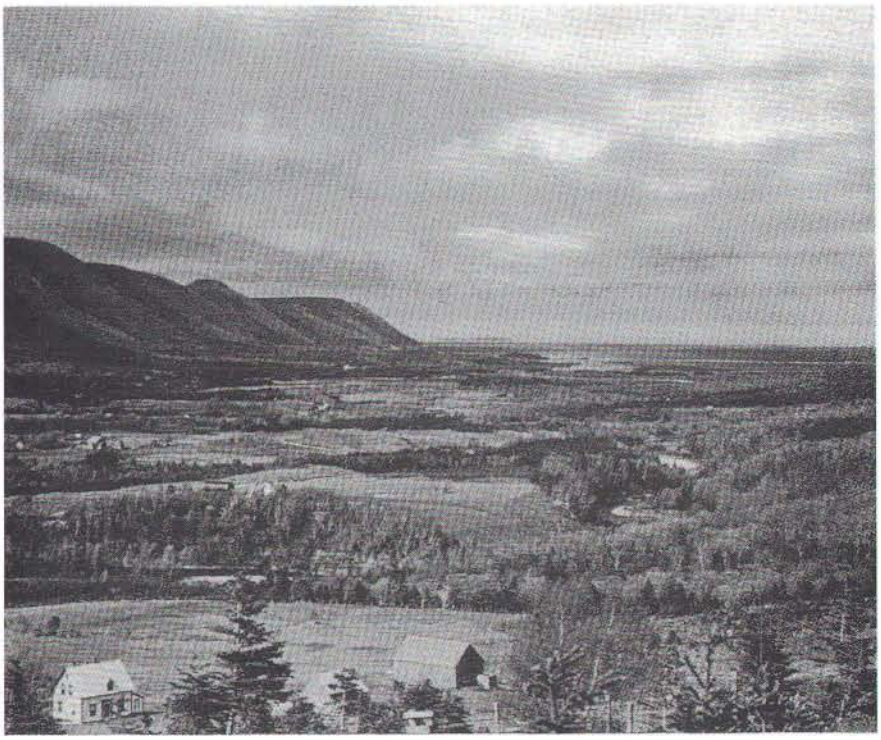
uranium, thorium and one isotope of potassium generates a large but unknown supply of heat that must have slowed, although it did not necessarily stop, the cooling of the earth.

The rigid earth now appeared to be less rigid. It became possible to explain the knowledge, already a century old, that great continental ice sheets had depressed the earth's crust, just as the loads of ice that cover Greenland and Antarctica depress the crust in those regions today. Observation showed that central Scandinavia and northern Canada, which had been covered with glacial ice until it melted 11,000 years ago, were still rising at the rate of about a centimeter a year. Calculations of the viscosity of the interior based on these studies led to the realization that the earth as a whole behaves as though a cool and brittle upper layer, perhaps 100 kilometers thick, rests on a hot and plastic interior. All the large topographical features—continents, ocean basins, mountain ranges and even individual volcanoes—slowly seek a rough hydrostatic equilibrium with one another on the exterior. Precise local measurements of gravity showed that the reason some features remain higher than others is that they have deeper, lighter roots than those that are low. In this way, continents were seen to float like great tabular icebergs on a frozen sea.

Everyone could agree on the fact that in response to vertical forces the outer crustal layer moved up and down, causing flow in the interior. The crux of the argument between the proponents of fixed and of drifting continents became the question of whether the outer crust must remain rigid under horizontal forces or whether it could respond to such forces by slow lateral moves.

Suggestions that the continents might have moved had been advanced on various grounds for centuries. The remarkable jigsaw-puzzle fit of the Atlantic coasts of Africa and South America provoked the imagination of explorers almost as soon as the continental outlines appeared opposite each other on the world map. In the late 19th century, geologists of the Southern Hemisphere were moved to push the continents of that hemisphere together in one or another combination in order to explain the parallel formations they found, and by the turn of the century the Austrian geologist Eduard Suess had reassembled them all

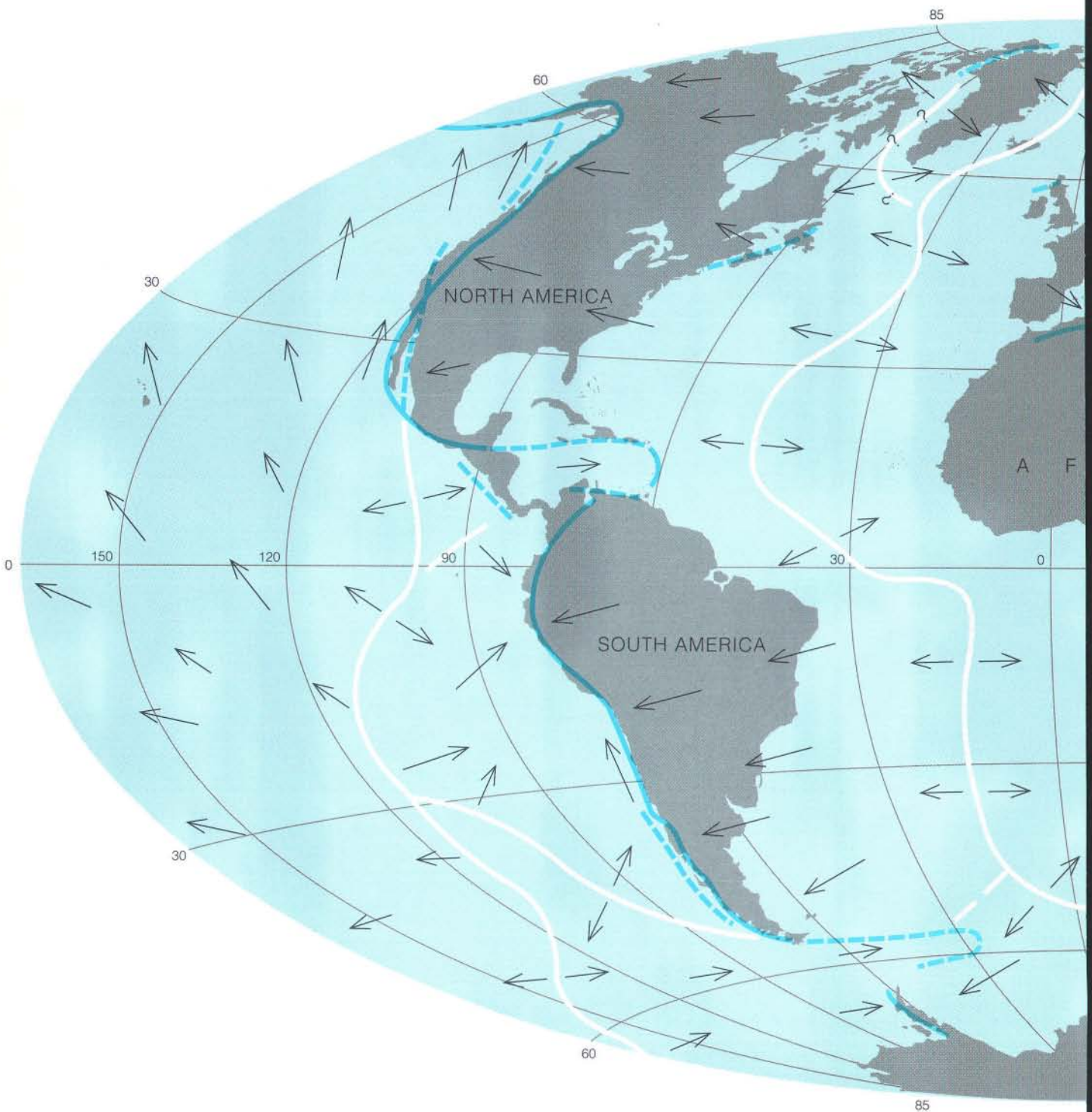
GREAT GLEN FAULT in Scotland is named for a valley resulting from erosion along the line of the fault. About 350 million years ago the northern part of Scotland was slowly moved some 60 miles to the southwest along this line.



ASPY FAULT in northern Nova Scotia is marked by several cliffs like the one seen here. The fault is part of the Cabot fault system extending from Boston to Newfoundland and may represent an extension of the Great Glen fault.

in a single giant landmass that he called Gondwanaland (after Gondwana, a key geologic province in east central India).

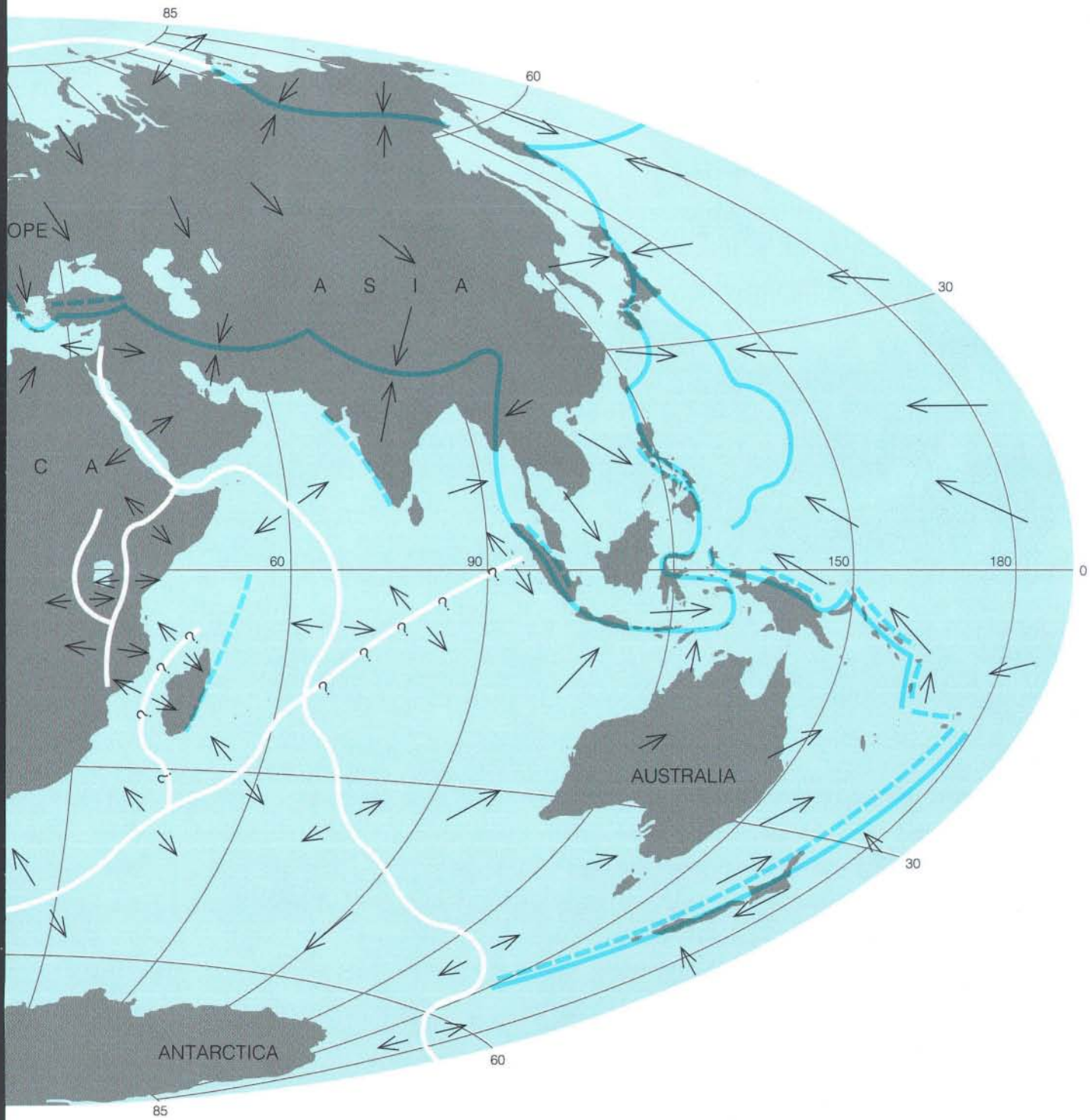
The first comprehensive theory of continental drift was put forward by the German meteorologist Alfred Wegener in 1912. He argued that if the earth could flow vertically in response to vertical forces, it could also flow laterally. In support of a different primeval arrangement of landmasses, he was able to point to an astonishing number of close affinities of fossils, rocks and structures on opposite sides of the Atlantic that, he suggested, ran evenly across, like lines of print when the ragged edges of two pieces of a torn newspaper are fitted together again.



CONVECTION CURRENTS in the earth's mantle may move blocks of crustal material with different effects. Continental mountain chains and island arcs could form where currents sink and blocks meet; mid-ocean ridges, where currents rise

According to Wegener, all the continents had been joined in a single supercontinent about 200 million years ago, with the Western Hemisphere continents moved eastward and butted against the western shores of Europe and Africa and with the Southern Hemisphere continents nestled together on the southern flank of this "Pangaea," as Wegener named it. Under the action of forces associated with the rotation of the earth, the continents had broken apart, opening up the Atlantic and Indian oceans.

Between 1920 and 1930 Wegener's hypothesis excited great controversy. Physicists found the mechanism he had proposed inadequate and expressed

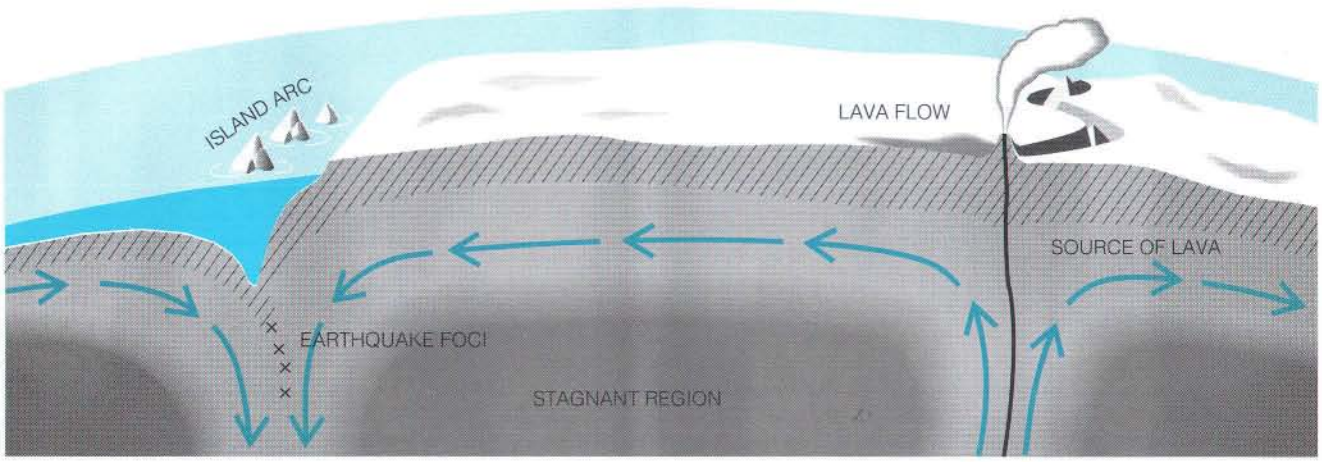


and blocks are torn apart. Arrows indicate directions of horizontal flow of currents. Solid colored lines represent mountain chains and island arcs; heavy white lines, the worldwide system of mid-ocean ridges; and broken colored lines, faults.

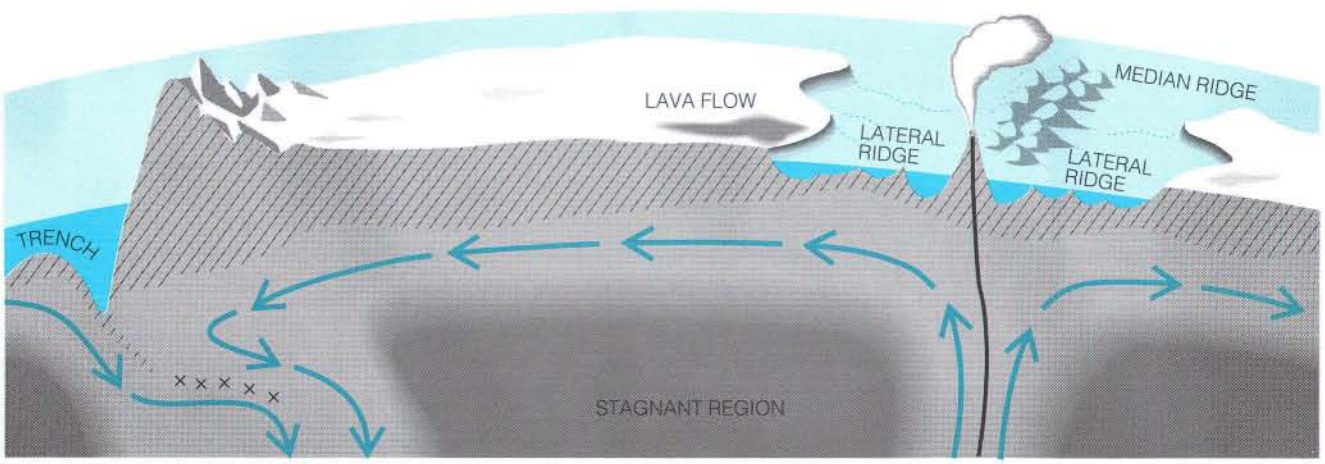
doubt that the continents could move laterally in any case. Geologists showed that some of Wegener's suggestions for reassembling the continents into a single continent were certainly wrong and that drift was unnecessary to explain the coincidences of geology in many areas. They could not, however, dispute the validity of most of the transatlantic connections. Indeed, more such connections have been steadily added.

It was the discovery of one of these connections that prompted my own recent inquiries into the subject of continental drift. A huge fault of great age bisects Scotland along the Great Glen in the Caledonian Mountains. On the western side of the Atlantic, I was able to show, a string of well-known faults of the same great age connect up into another huge fault, the "Cabot fault" extending from Boston to northern Newfoundland. These two great faults are much older than the submarine ridge and rift recently discovered on the floor of the mid-Atlantic and shown to be a young formation. The two faults would be one if Wegener's reconstruction or something like it were correct. Wegener also thought that Greenland (where he died in 1930) and Ellesmere Island in the Canadian Arctic had been torn apart by a great lateral displacement along the Robeson Channel. The Geological Survey of Canada has since discovered that the Canadian coast is faulted there.

Many geologists of the Southern Hemisphere, led by Alex L. Du Toit of South Africa, welcomed Wegener's views. They sought to explain the mounting evidence that an ice age of 200 million years ago had spread a glacier over the now scattered continents of the Southern Hemisphere. At the same time, ac-



EFFECTS OF CONVECTION CURRENTS, schematized in the two illustrations on this page, provide one possible means of accounting for the formation of median ridges, lateral ridges, mountain ranges and earthquake belts. Rising and separating currents (*arrows at right*) could break the crustal rock and pull it apart; the rift would be filled by altered mantle material and lava flows, forming a median ridge. Sinking currents (*left*) could pull the ocean floor down.



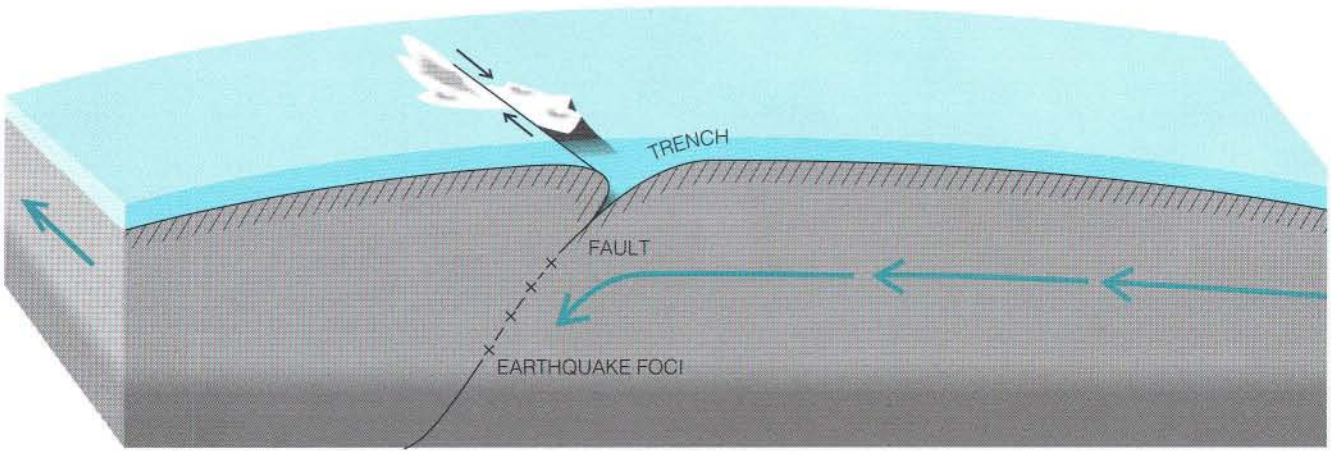
DRIFTING CONTINENT may be "piled up," where it meets sinking currents, to form mountains like those of the Andes (*left*). Since continents are lighter than the mantle material of the ocean floor, they cannot sink but tend to be pushed over sinking currents, marked by deep earthquakes. Volcanoes continue to form over rising currents, but drift may carry these volcanic piles away to either side of the ridge. Separated from their source, the inactive cones form one or two lateral ridges.

cording to the geologic record, the great coal deposits of the Northern Hemisphere were being formed in tropical forests as far north as Spitsbergen. To resolve this climatic paradox, Du Toit proposed a different reconstruction of the continent. He brought the southern continents together at the South Pole and the northern coal forests toward the Equator. Later, he thought, the southern continent had broken up, and its component subcontinents had drifted northward.

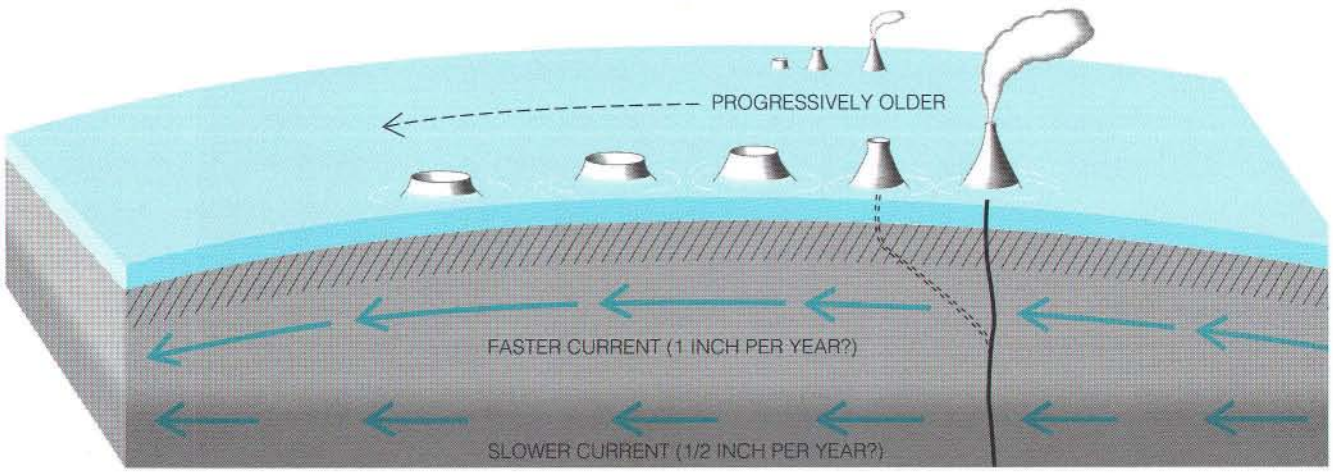
The compelling evidence for the existence of a Gondwanaland during the Mesozoic era—the Age of Reptiles—has been reinforced by the findings made in Antarctica since the intensive study of that continent began in 1955. The ice-free outcrops on the continent, although few, not only show the record of the earlier ice age that gripped the rest of the landmasses in the Southern Hemisphere but also bear deposits of a low-grade coal laid down in a still earlier age of verdure that covered all the same landmasses with the peculiar big-leafed *Glossopteris* flora found in their coal beds as well.

Many suggestions have been made as to how to create and destroy the land bridges needed to explain the biological evidence without moving the continents. Some involve isthmuses, and some involve whole continents that have subsided below the surface of the ocean. But the chemistry and density of continents and ocean floors are now known to be so different that it seems even more difficult today to raise and lower ocean floors than it is to cause continents to migrate.

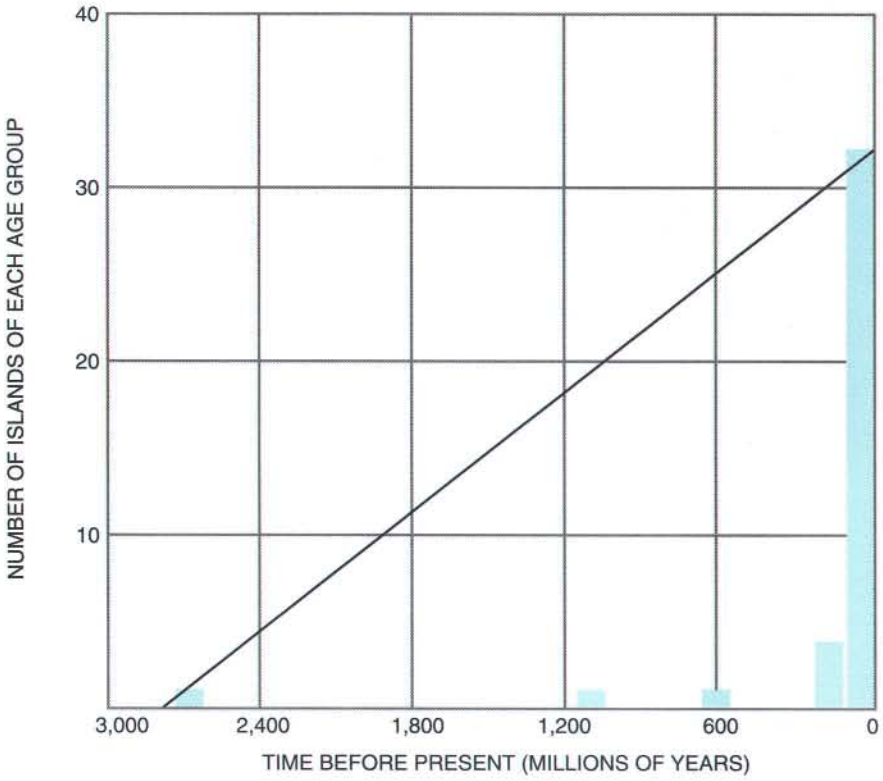
One of the first leads to a mechanism that would move continents came more than 30 years ago from the ex-



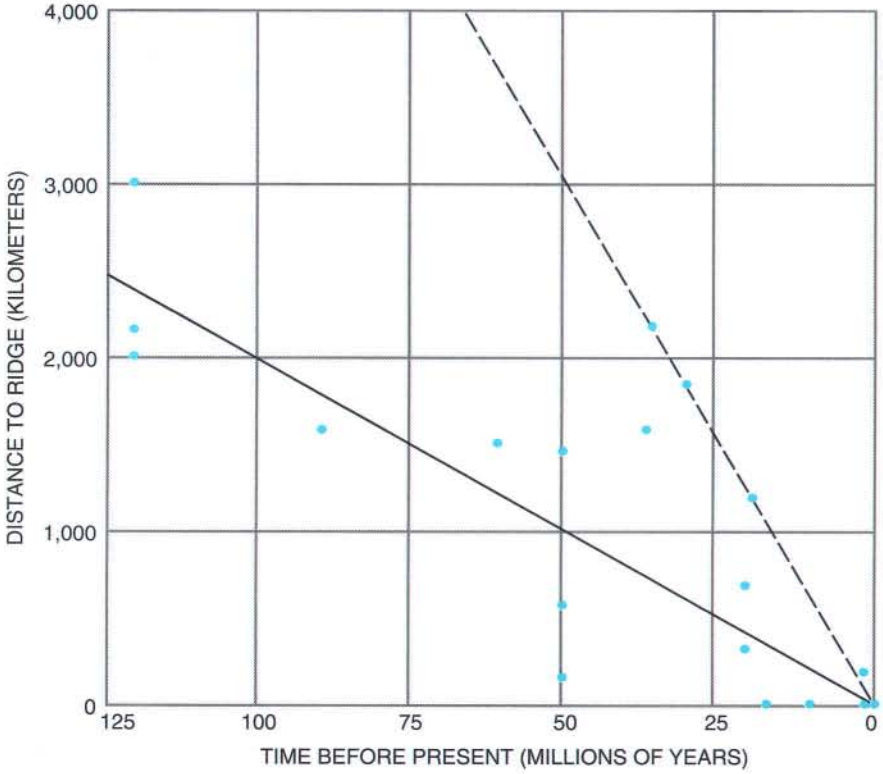
TWO CONVECTION CURRENTS perpendicular to each other suggest a mechanism for producing large horizontal faults such as the one that has offset western New Zealand 300 miles northward. The two convection currents would produce a fault. One current would be forced downward, producing a trench and earthquakes along the sloping surface. Continued flow of the second current would result in a sliding motion along the plane of the fault, shearing the island in two.



VOLCANIC ISLAND CHAINS like the Hawaiian Islands must have originated in a process slightly different from that which formed pairs of lateral ridges. The source of lava flow does not lie on a mid-ocean ridge; it is considered that the source may be deep (100 miles or more) in the slower moving part of convection currents. The differential motion carries old volcanoes away from the source, while new volcanoes form over the source.



FREQUENCY DIAGRAM shows the age distribution of about 40 islands (in main ocean basins) dated older than "recent." The diagonal line shows the corresponding curve for continental rock ages over equivalent areas.



DISTANCE FROM MID-OCEAN RIDGE of some islands in Atlantic and Indian oceans is plotted against age. If all originated over the ridge, their average rate of motion has been two centimeters a year (*solid line*); maximum rate, six (*broken line*).

tension to the ocean floor of the sensitive techniques of gravimetry that had established the rule of hydrostatic equilibrium, or isostasy, ashore. The Dutch geophysicist Felix A. Vening Meinesz demonstrated that a submerged submarine would provide a sufficiently stable platform to allow the use of a gravimeter at sea. Over the abyssal trenches in the seafloor that are associated with the island arcs of Indonesia and the western side of the Pacific, he found some of the largest deficiencies in gravity ever recorded. It was clear that isostasy does not hold in the trenches. Some force at work there pulls the crust into the depths of the trenches more strongly than the pull of gravity does.

Arthur Holmes of the University of Edinburgh and D. T. Griggs, now at the University of California at Los Angeles, were stimulated by these observations to reexamine and restate in modern terms an old idea of geophysics: that the interior of the earth is in a state of extremely sluggish thermal convection, turning over the way water does when it is heated in a pan. They showed that convection currents were necessary to account in full for the transfer of heat flowing from the earth's interior through the poorly conductive material of the mantle: the region that lies between the core and the crust. The trenches, they said, mark the places where currents in the mantle descend again into the interior of the earth, pulling down the ocean floor.

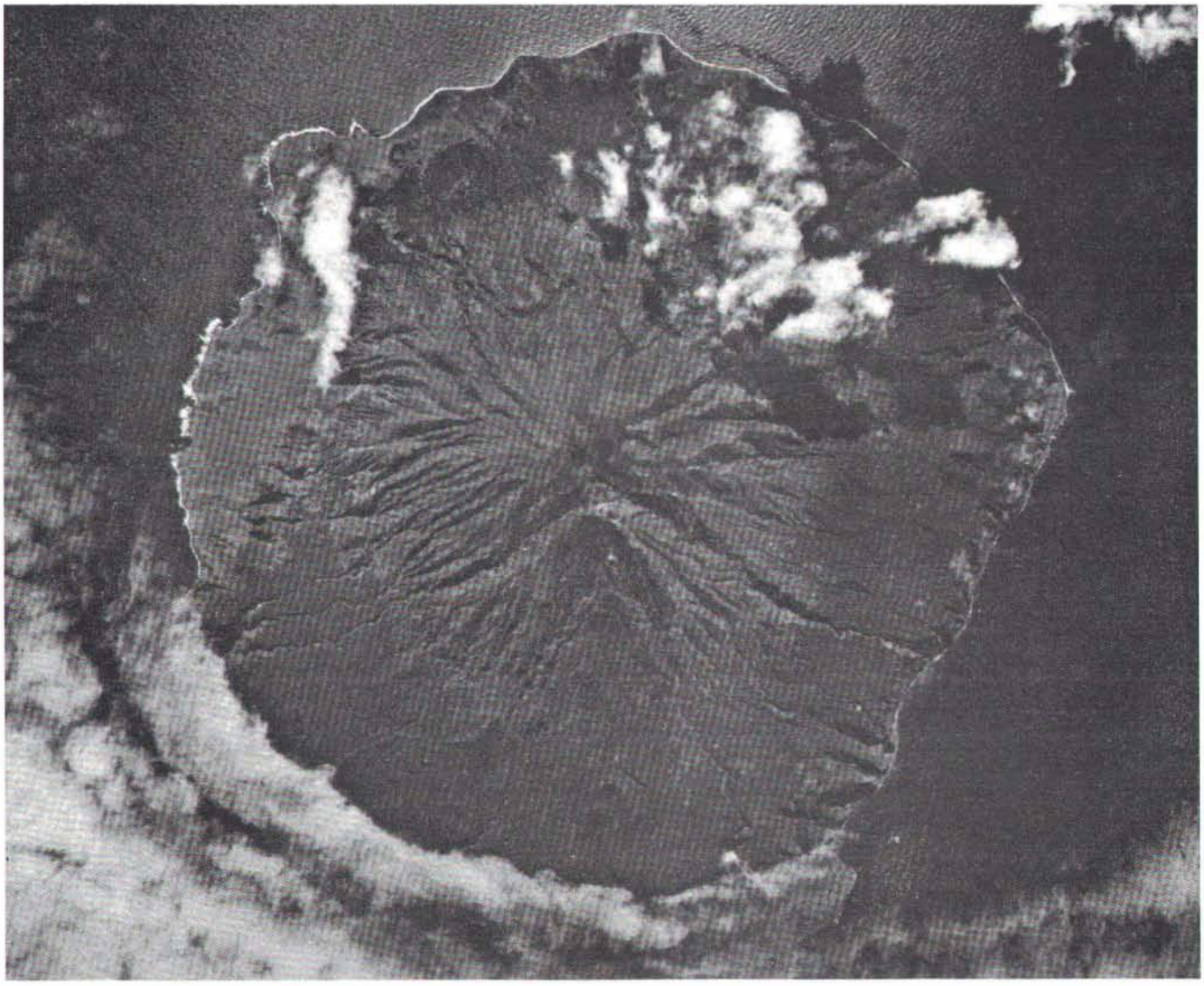
Convection currents in the mantle now play the leading role in every discussion of the large-scale and long-term processes that go on in the earth. It is true that the evidence for their existence is indirect; the currents flow too deep in the earth and too slowly—a few centimeters a year—for direct observation. Nonetheless, their presence is supported by an increasing body of independently established evidence and by a more rigorous statement of the theory of their behavior. Recently, for example, S. K. Runcorn of Durham University has shown that to stop convection the mantle material would have to be 10,000 times more viscous than the rate of postglacial recoil indicates. It is, therefore, highly probable that convection currents are flowing in the earth.

Perhaps the strongest confirmation has come with the discovery of the regions where these currents appear to ascend toward the earth's surface. This

is the major discovery of the recent period of extraordinary progress in the exploration of the ocean bottom, and it involves a feature of the earth's topography as grand in scale as the continents themselves. Across the floors of all the oceans, for a distance of 40,000 miles, there runs a continuous system of ridges. Over long stretches, as in the mid-Atlantic, the ridge is faulted and rifted under the tension of forces acting at right angles to the axis of the ridge. Measurements first undertaken by Sir Edward Bullard of the University of Cambridge show that the flow of heat is unusually great along these ridges, exceeding by two to eight times the average flow of a millionth of a calorie per square centimeter per second observed on the continents and elsewhere on the ocean floor. Such measurements also show that the flow of heat in the trenches, as in the Acapulco Trench off the Pacific coast of Central America, falls to as little as a tenth of the average.

Most oceanographers now agree that the ridges form where convection currents rise in the earth's mantle and that the trenches are pulled down by the descent of these currents into the mantle. The possibility of lateral movement of the currents in between is supported by evidence for a slightly plastic layer—called the asthenosphere—below the brittle shell of the earth. Seismic observations show that the speed of sound in this layer suddenly becomes slower, indicating that the rock is less dense, hotter and more plastic. These observations have also yielded evidence that the asthenosphere is a few hundred kilometers thick, somewhat thicker than the crust, and that below it the viscosity increases again.

Here, then, is a mechanism, in harmony with physical theory and much geologic and geophysical observation, that provides a means for disrupting and moving continents. It is easy to believe that where the convection currents rise and separate, the surface rocks are broken by tension and pulled apart, the rift being filled by the altered top of the mantle

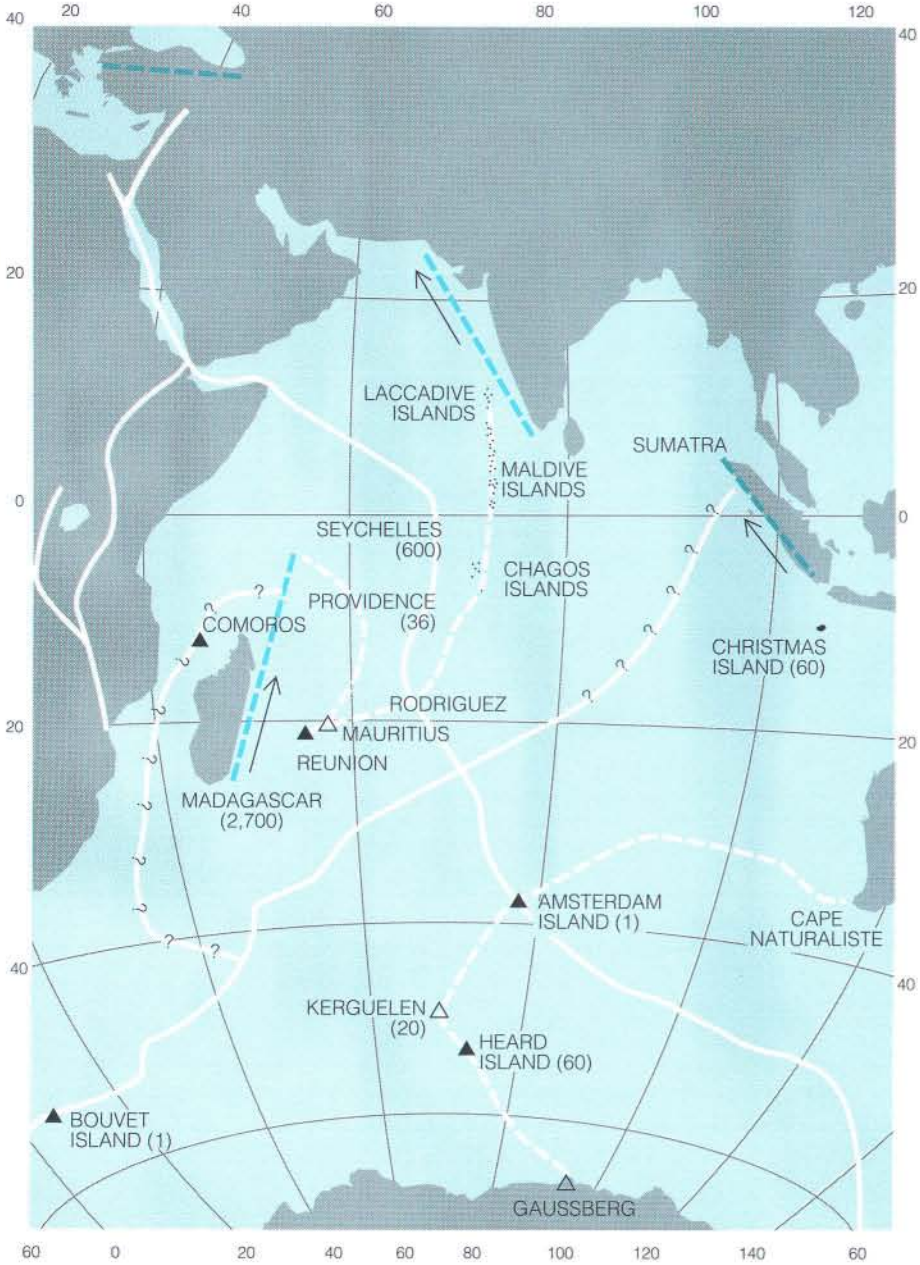


TRISTAN DA CUNHA ISLAND in the South Atlantic lies on the Mid-Atlantic Ridge. At center are the lava beds and partially filled crater of the main cone, which has not erupted for several centuries. Along the perimeter of the island, secondary cones are just discernible, as is the settlement on the island's northeastern promontory (*upper left*). Several months after this aerial photograph was made in 1961, a volcanic eruption took place about 300 yards east (*to right*) of the settlement.

and by the flow of basalt lavas. In contrast to earlier theories of continental drift that required the continents to be driven through the crust like ships through a frozen sea, this mechanism conveys them passively by the lateral movement of the crust from the source of a convection current to its sink. The continents, having been built up by the accumulation of lighter and more siliceous materials brought up from below, are not dragged down at the trenches where the currents descend but pile up there in mountains. The ocean floor, being essentially altered mantle, can be carried downward; such sediments as have accumulated in the trenches descend also and, by complicated processes, may add new mountains to the continents. Since the material near the surface is chilled and brittle, it fractures, causing earthquakes until it is heated by its descent.

From the physical point of view, the convection cells in the mantle that drive these currents can assume a variety of sizes and configurations, starting up and slowing down from time to time, expanding and contracting. The flow of the currents on the world map may therefore follow a single pattern for a time, but the pattern should also change occasionally because of changes in the output and transfer of heat from within. It is thus possible to explain the periodicity of mountain-building, the random and asymmetric distribution of the continents and the abrupt breakup of an ancient continent.

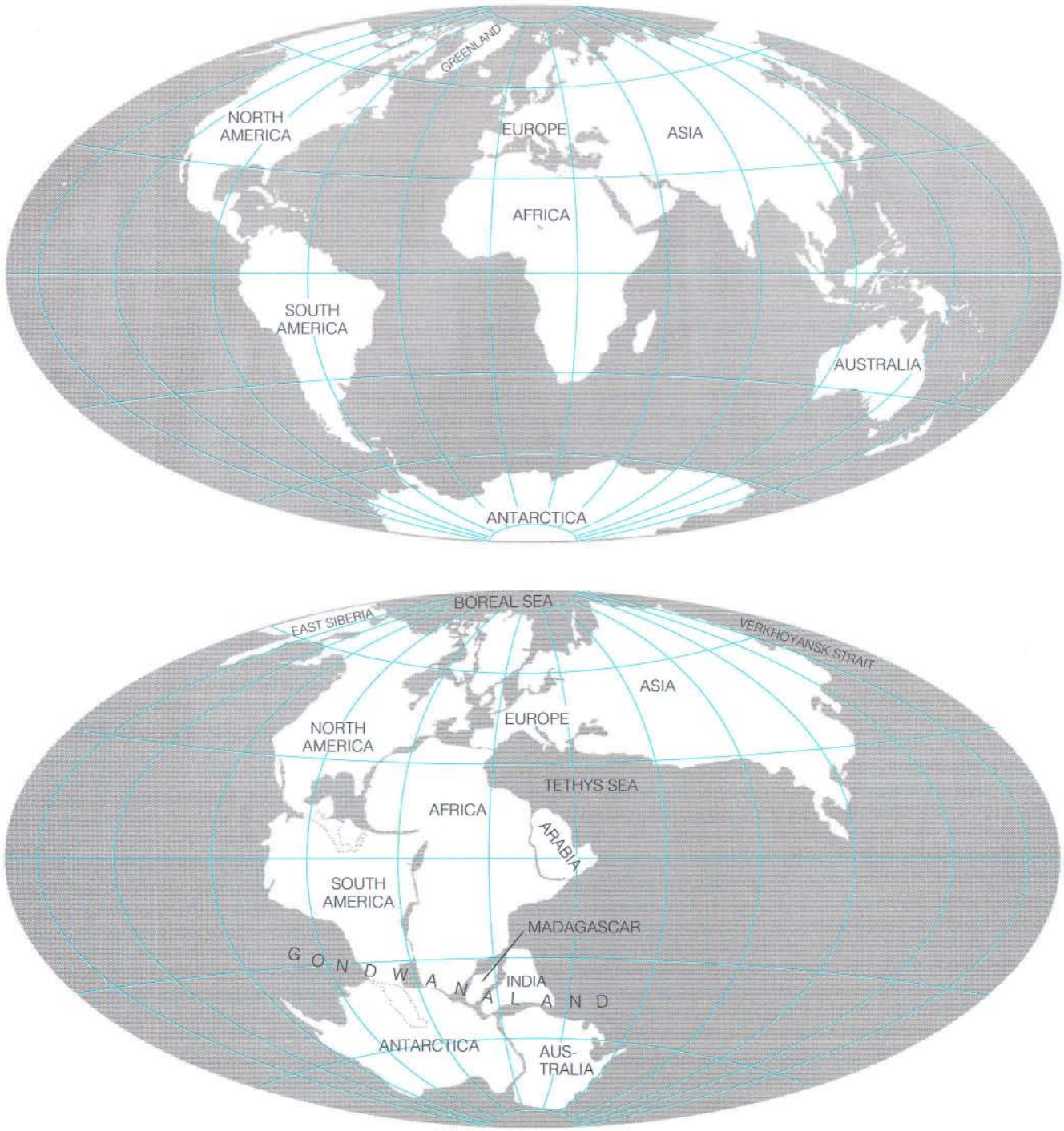
Some geophysicists consider that isostatic processes set up by gravitational forces may suffice to cause the outer shell to fracture and to slip laterally over the plastic layer of the asthenosphere. This mechanism would not require the intervention of convection currents. Both mechanisms could explain large horizontal displacements of the crust.



INDIAN OCEAN possibly formed as the result of four continents drifting apart. If so, four median ridges would have formed midway between continents, with pairs of lateral ridges connecting them. The heavy white lines show three known median ridges; there is evidence for one running to Sumatra. The broken white lines are lateral ridges; broken colored lines, faults; open triangles, inactive volcanoes. The numbers give ages in millions of years.

Fresh evidence that such great movements have indeed been taking place has been provided by two lines of study in the field of terrestrial magnetism. On the one hand, surveys of the earth's magnetic field off the coast of California show a pattern of local anomalies in the ocean floor running parallel to the axis of a currently inactive oceanic ridge that underlies the edge of the continent. The pattern bears a persuasive resemblance to the "photoelastic" strain patterns revealed by polarized light in plastics placed under stress. More important, the pattern shows that the ocean floor is faulted at right angles to the axis of the ridge, with great slabs of the crust displaced laterally to the west by as much as 750 miles. These are apparently ancient and inactive fractures; now the active faults run northwesterly, as is indicated by the earthquakes along California's San Andreas fault.

Evidence of a more general nature in favor of continental drift comes from the studies of the "remanent" magnetism of the rocks, to which Runcorn, P.M.S. Blackett of the University of London and Emil Thellier of the University of Paris have made significant contributions. Their investigations have demonstrated that rocks can be weakly magnetized at the time of formation—during cooling in the case of lavas and during deposition in the case of sediments—and that their polarity is aligned with the direction of the earth's magnetic field at the place and time of



SINGLE SUPERCONTINENT, presumed to have existed some 150 million years ago, would have resembled that depicted in the map at bottom. A present-day map appears at top. The distortion of the continents is a result of the projection used.

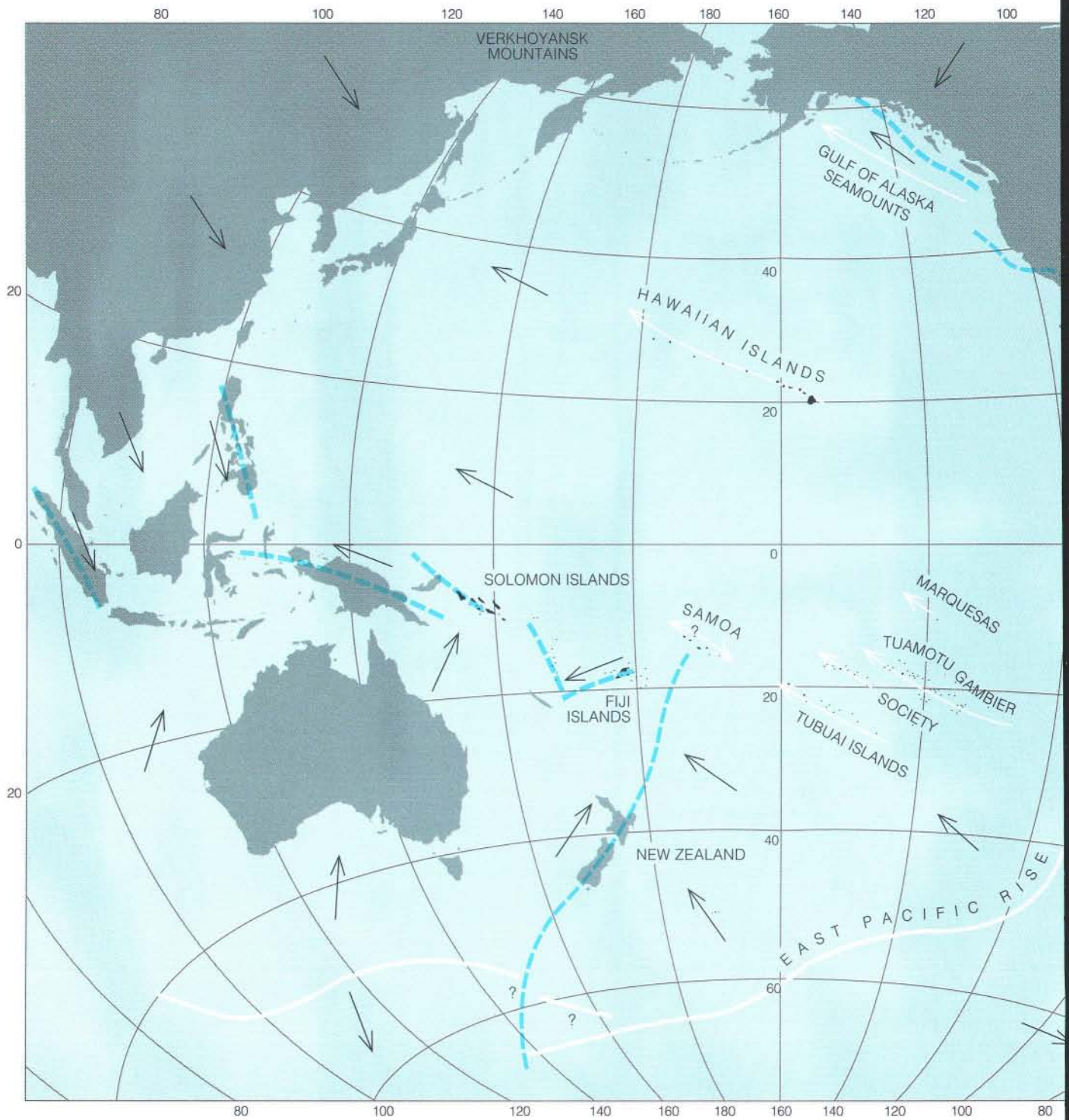
their formation. The present orientation of the rocks of various ages on the continents indicates that they must have been formed in different latitudes. The rocks of any one continent show consistent trends in change of orientation with age; those from other continents show different shifts. Continental drift offers the only explanation of these findings that has withstood analysis.

Some physicists and biologists are now prepared to accept continental drift, but many geologists still have no use for the hypothesis. This is to be expected. Continents are so large that much geology would be the same whether drift had occurred or not. It is the geology of the ocean floors that promises to settle the question in time, but the real study of that two thirds of the earth's surface has just begun.

One test turns on the age of the ocean floor. If the continents have been fixed,

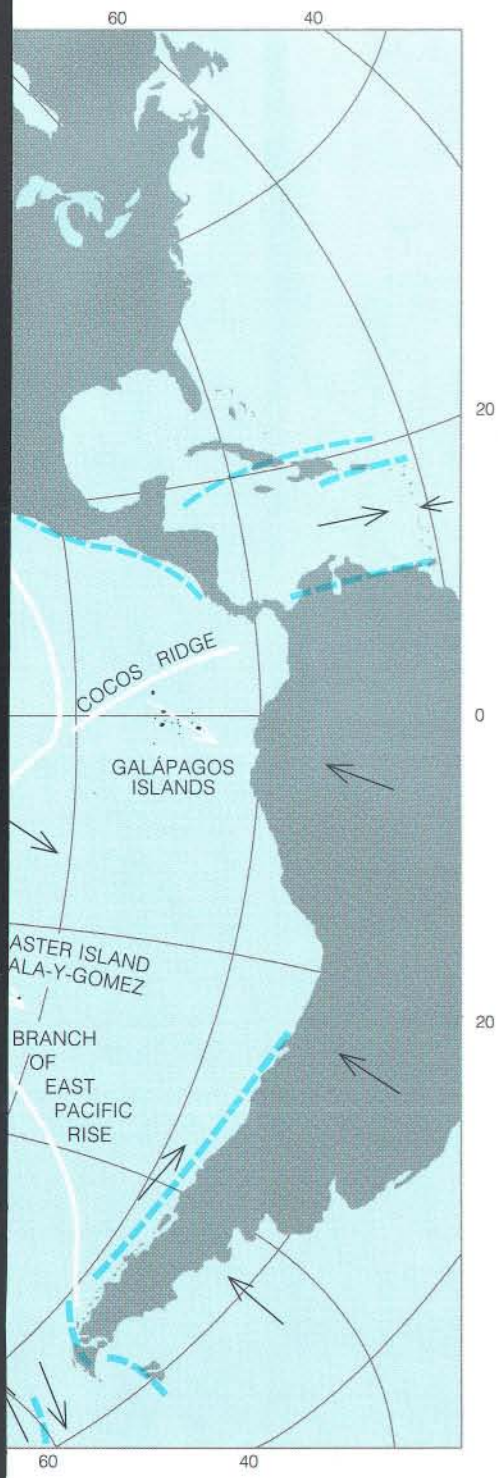
the ocean basins should all be as old as the continents. If drift has occurred, some regions of the ocean floor should be younger than the time of drift.

A survey of the scattered and by no means complete literature on the oceanic islands conducted by our group at the University of Toronto shows that of all the islands in the main ocean basins only about 40 have rocks that have been dated older than the Recent epoch. Only three of these—Madagascar and the Seychelles of the Indian Ocean and the Falklands of the South Atlantic—have very old rocks; all the



AGE OF PACIFIC ISLANDS appears to increase with increasing distance from the mid-ocean ridge. This is compatible with the idea that the eastern half of the Pacific Ocean has been spreading from the East Pacific Rise. Broken colored lines represent faults; the associated arrows indicate the direction of horizontal motion, where known, along the fault. Other arrows show the probable directions of convection flow. Island arcs of the kind represented by Japan develop

others are less than 150 million years old. If one regards the exceptions as fragments of the nearby continents, the youth of the other islands suggests either that the ocean basins are young or that islands are not representative samples of the rock of the ocean floor.



where the forces associated with such flow are directly opposed; great horizontal faults develop where these forces meet at right angles.

Significantly, it turns out that the age of the islands in the Atlantic Ocean tends to increase with their distance from the mid-ocean ridge. In this reckoning, one need not count the island arcs of the West Indies or the South Sandwich Islands, which belong to the Cordilleran system—that is, the spine of mountains running the entire length of North and South America—and so have a continental origin. At least six of the islands on the ridge or very close to it have on them active volcanoes that have had recent eruptions; the most recent was the eruption of Tristan da Cunha, which is located squarely on the ridge in the South Atlantic. Only two of the islands far from the ridge have active volcanoes. If the hot convection currents of the mantle rise under the mid-ocean ridge, it is easy to understand why the ridge is the locus of active volcanoes and earthquakes. The increase in age with distance from the ridge suggests that if the more distant islands had a volcanic origin on the ridge, lateral movement of the ocean floor has carried them away from the ridge. Their ages and distances from the ridge indicate movement at the rate of two to six centimeters a year on the average, in keeping with the estimated velocity of the convection currents.

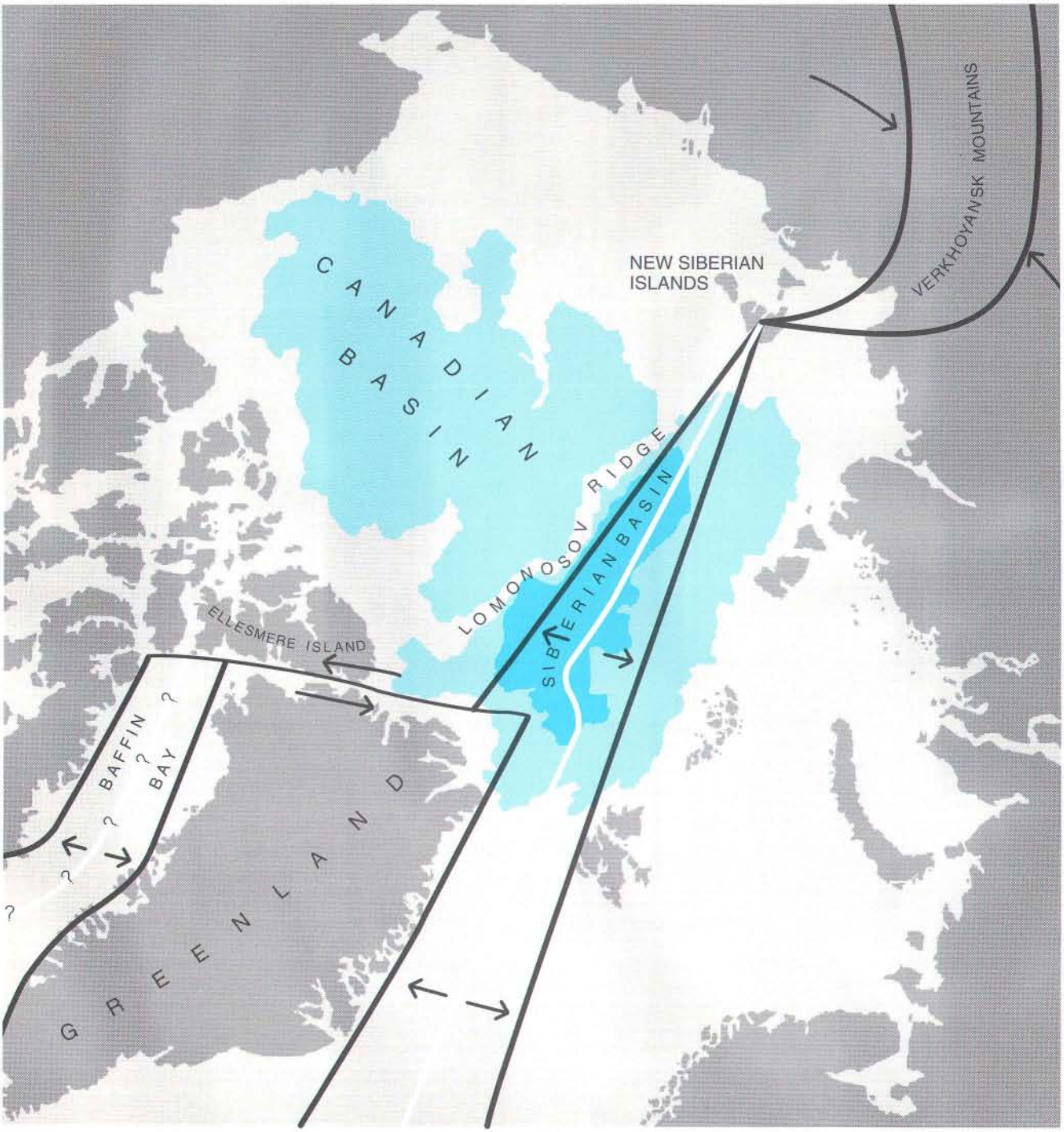
Of great significance in connection with the mechanism postulated here are the two lateral ridges that run east and west from Tristan da Cunha to Africa on the one hand and to South America on the other. It is reasonable to suppose that these ridges had their origin in a succession of volcanoes that erupted and grew into mountains on the site of the present volcano and were carried off east and west to form a row of progressively older, extinct and drowned volcanoes. There have been no earthquakes along the lateral ridges, and so they are distinctly different in character from the mid-ocean ridge. These ridges meet the continental margins at places that would fit together on the quite independent criterion of the match of their shorelines. One explanation of this coincidence is that the continents were indeed joined together and have moved apart, with the lateral ridges forming trails that record the motion. The two lateral ridges are roughly mirror images of each other, showing that the motion was uniform on each side. Another similar pair of ridges connects Iceland—where the mid-ocean ridge comes to the surface and where the great tension rift is visible in the Icelandic Graben—to Greenland and the shelf of the European continent.

We have therefore advanced two related hypotheses: first, where adjacent continents were once joined a median ridge should now lie between them; second, where such continents are connected by lateral ridges they were once butted together in such a manner that points marked by the shoreward ends of these ridges coincided. If this is correct, it provides a unique method for reassembling continents that have drifted apart. One of the major troubles with theories of drift has been that the possibilities are so numerous no such precise criterion existed for putting the poorly fitting jigsaw puzzle together.

Without doubt the most severe test of this dual hypothesis is presented by the Indian Ocean. Here four continents—Africa, India, Australia and Antarctica—may be assumed on geologic and paleomagnetic evidence to have drifted apart. The collision of India with the Asian landmass could have thrown up the Himalaya Mountains at their junction. These continents should accordingly be separated by four mid-ocean ridges. Three such ridges have already been well established by surveys of the Indian Ocean, and there is evidence for the existence of the fourth. In each quadrant marked off by the ridges, there is also, it happens, a lateral ridge! These submarine trails may be presumed to be records of the motion of the continents as they receded from one another. From Amsterdam Island one of these lateral ridges runs through Kerguelen Island to Gaussberg Mountain on the coast of Antarctica; a mirror image of this ridge runs from Amsterdam Island to Cape Naturaliste in Australia. The corresponding ridges connecting Africa and India are distorted by lateral faults running along the coasts of Madagascar and India. Thus in each quadrant there exists a lateral ridge to show how points in Madagascar, India, Australia and Antarctica once lay close together. What is remarkable is not that there is some irregularity in the present configuration of these ridges but that the floor of the Indian Ocean should show such a symmetric pattern.

The mid-ocean ridge separating Australia from Antarctica has been traced by Henry W. Menard of the Scripps Institution of Oceanography across the eastern Pacific to connect with the great East Pacific Rise. From the topography of the Pacific floor it can be de-

duced that this ridge once extended through the rise marked by Cocos Island off Central America and formed the rifted ridge that moved North and South America apart. Another branch of this ridge, running across the southern latitudes, suggests the cause of the separation of South America from Antarctica. The oceanic islands in this broad region of the Pacific form lines that extend at right angles down the flanks of the East Pacific Rise; geologists long ago established that these islands grow progressively older with distance from the top of the rise. Unlike the rest of the continuous belt of mid-ocean ridges to which it is connected, the East Pacific Rise tends to run along the margins of the Pacific Ocean; it has rifted an older ocean apart rather than a continent. The floor



RIFTING OF SUPERCONTINENT to form the Atlantic Ocean could have produced the Verkhoyansk Mountains in eastern Siberia. The rift spread more widely to the south. The opening of the Atlantic Ocean and Baffin Bay separated Greenland from both North America and Europe. The continents were rotated slightly about a fulcrum near the New Siberian Islands. The resulting compression and uplift would create a mountain range. Opposing arrows mark the Wegener fault.

of the western Pacific is believed to be a remnant of that older floor.

There are therefore enough connections to draw all the continents together, reversing the trends of motion indicated by the mid-ocean ridges and using the continental ends of pairs of lateral ridges as the means of matching the coastlines together. The ages of the islands and of the coastal formations suggest that about 150 million years ago, in mid-Mesozoic time, all the continents were joined in one landmass and that there was only one great ocean. The supercontinent that emerges from this reconstruction is not the same as those proposed by Wegener, Du Toit and other geologists, although all have features in common. The widespread desert conditions of the mid-Mesozoic may have been a consequence of the unusual circumstance that produced a single continent and a single ocean at that time. Since its approximate location with respect to latitude is known, along with the location of its major mountain systems, the climate in various regions might be reconstructed and compared with geologic evidence.

It is not suggested that this continent was primeval. That it was in fact assembled from still older fragments is suggested by two junction lines: the ancient mountain chain of the Urals and the chain formed by the union of the Appalachian, Caledonian and Scandinavian mountains may have been thrown up in the collisions of older continental blocks. Before that there had presumably been a long history of periodic assembly and disassembly of continents and fracturing and spreading of ocean floors, as convection cells in the mantle proceeded to turn over in different configurations. At the present writing it is impossible even to speculate about the details.

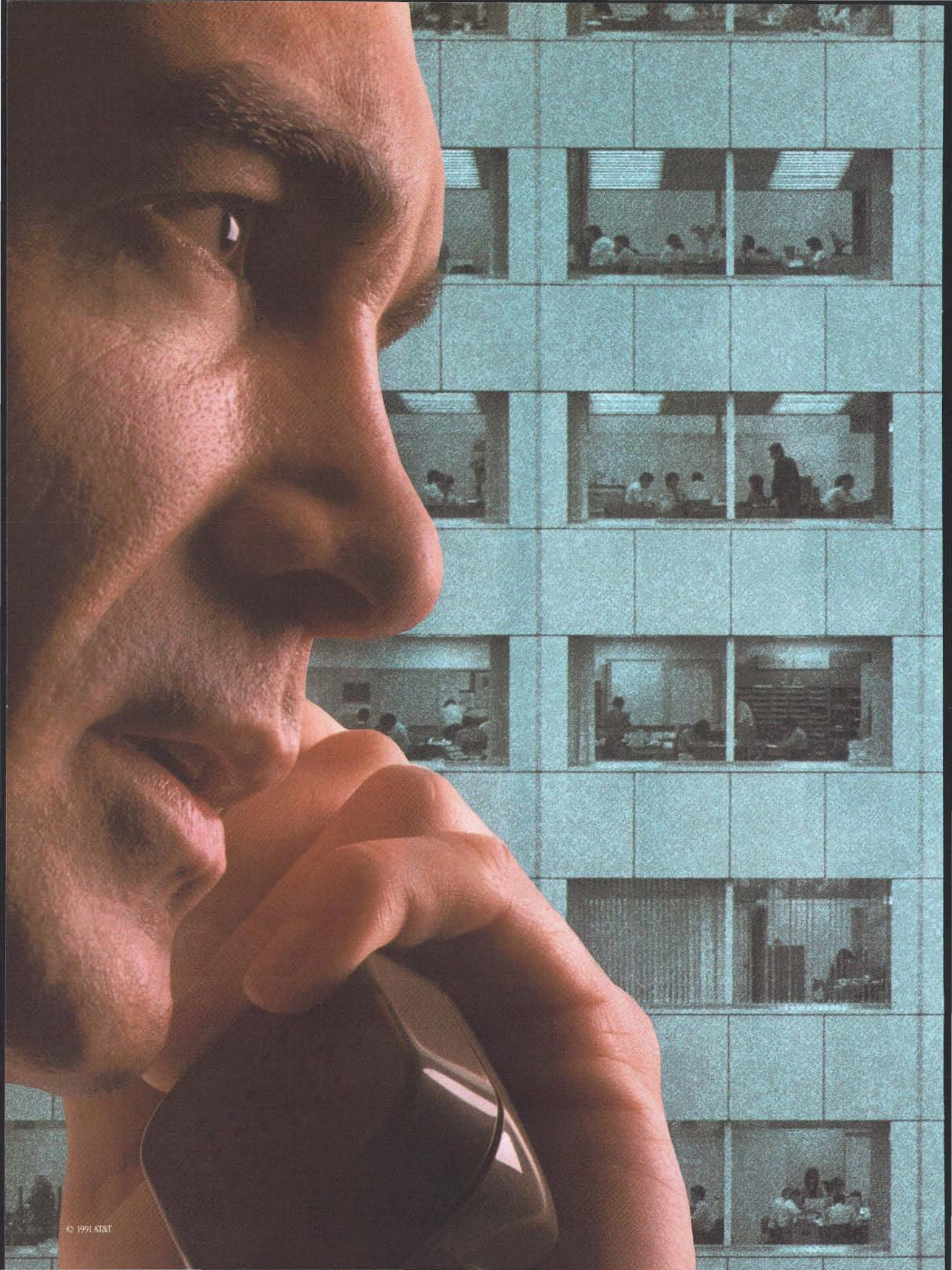
If it can be assumed that the proposed Mesozoic continent did exist and spread apart, geology provides some guide to the history of its fragmentation. The present system of convection currents has apparently been constant in general configuration ever since the Mesozoic, but not all parts of it have been equally active all of that time. Shortly before the start of the Cretaceous period, about 120 million years ago, the continent developed a rift that opened up to form the Atlantic Ocean. The rift spread more widely in the south, with the result that the continents must have rotated slightly about a fulcrum near the New Siberian Islands. Soviet geologists have found that the compression and uplift that raised the Verkhoyansk Mountains across eastern Siberia began at about that time. To the south a continuation of the rifting separated Africa from Antarctica and spread diagonally across the Indian Ocean, opening the northeasterly rift. Africa and India were thus moved northward, away from the still intact Australian-Antarctic landmass.

It seems reasonable to suggest, particularly from the geology of the Verkhoyansk Mountains and of Iceland, that at the start of Tertiary time, about 60 million years ago, this convection system became less active and that rifting started up elsewhere. A new rift opened up along the other, northwesterly, diagonal of the Indian Ocean, separating Africa from India and Australia and separating Australia from Antarctica. With the collision of the Indian subcontinent against the southern shelf of the Asiatic landmass, the uplift of the Himalaya Mountains began. The proposed succession of activity in the two main ridges of the Indian Ocean would explain why India has moved twice as far north with relation to Antarctica as Australia or Africa has and why the older northeast ridge is now a somewhat indistinct feature of the ocean floor. The younger rift in the Indian Ocean seems to have extended along the East Pacific Rise and Cocos Ridge to cross the Caribbean. A branch also passed south of South America. As these median ridges have continued to widen, they have been forced by this growth to migrate northward, forming great shears or faults off the coast of Chile and through California. Indeed, a case can be made out for the idea that every mid-ocean ridge normally ends at a great fault or at a pivot point, as in the New Siberian Islands.

A few million years ago activity in this system decreased, allowing the North and South American continents to be joined by the Isthmus of Panama. The Atlantic rift now became more active again, producing renewed uplift in the Verkhoyansk Mountains and active volcanoes in Iceland and the five other still active volcanic islands down the Atlantic. Again the pattern of rifting in the Indian Ocean was altered. The distribution of recent earthquakes shows that the greatest activity extends along the western half of each diagonal ridge from the South Atlantic to the entrance of the Red Sea and thence by two arms along the rift valley of the Jordan River and through the African rift valleys, where the breakup of a continent has apparently begun.

The currently expanding rifts run mostly north and south or northeasterly, so that dominant easterly and westerly compression of the outer crust is absorbed by overthrusting and sinking of the crust along the eastern and western sides of the "ring of fire" around the Pacific. For this reason East Asia, Oceania and the Andes are the most active regions of the world. The westward-driving pressure of the South Atlantic portion of the Mid-Atlantic Ridge has forced the continental block of South America against and over the downward-plunging oceanic trench along its Pacific coast. The northwest-trending currents below the Pacific floor have pulled down trenches under the eight island arcs around the western and northern Pacific from the Philippines north to the Aleutians. Even at the surface of the Pacific, the direction of the subcrustal movement is indicated by the strike of several parallel chains of volcanic islands, such as the Hawaiians, which may be thought to have risen like bubbles in a stream from the slower moving deep interior. These chains of volcanic islands run parallel with the seismically active shearing faults that border each side of the Pacific, along the coast of North America and from Samoa to the Philippines. The compression exerted by the mid-ocean ridge through the southern seas is absorbed, with less seismic activity, along a line from New Zealand, through Indonesia and the Himalaya highlands to the European Alps. In all cases, the angle at which the loci of deep-focus earthquakes dip into the earth seems to follow the direction of subsurface flow—eastward and downward, for example, under the Pacific coast of South America, and westward and downward under the island arcs on the opposite side of the Pacific.

The theory I have outlined may be highly speculative, but it is indicative of current trends in thought about the earth's behavior. The older theories of the earth's history and behavior have proved inadequate to meet the new findings, particularly those from investigations of terrestrial magnetism and oceanography. It is a fact in favor of the specific details suggested here that they fit observations and are precise enough to be tested.


© 1991 AT&T

Premises, Premises.

Or, How AT&T SYSTIMAX PDS Gives More Promise To Your Premises.

AT&T SYSTIMAX® Premises Distribution System gives you the capability to access voice, data, video. Any kind of information. Anywhere in your building. Through a single comprehensive building wiring system. AT&T SYSTIMAX PDS is a completely modular system that combines hardware, copper wire, connectors, and fiber optics; with the only five year product warranty in the business. AT&T SYSTIMAX PDS enables information to travel from building to building. Office to office. Floor to floor. Platform to platform. Regardless of differences in operating systems. Or networking protocols. A truly open architecture. For real multivendor compatibility. With AT&T SYSTIMAX PDS, as your business grows, your system grows, too. And that can give any business more promise. Call AT&T Network Systems at **1 800 344-0223, ext. 1026** to see how it can work for you.

AT&T Network Systems And Bell Laboratories. Technologies For The Real World.

AT&T
Network Systems

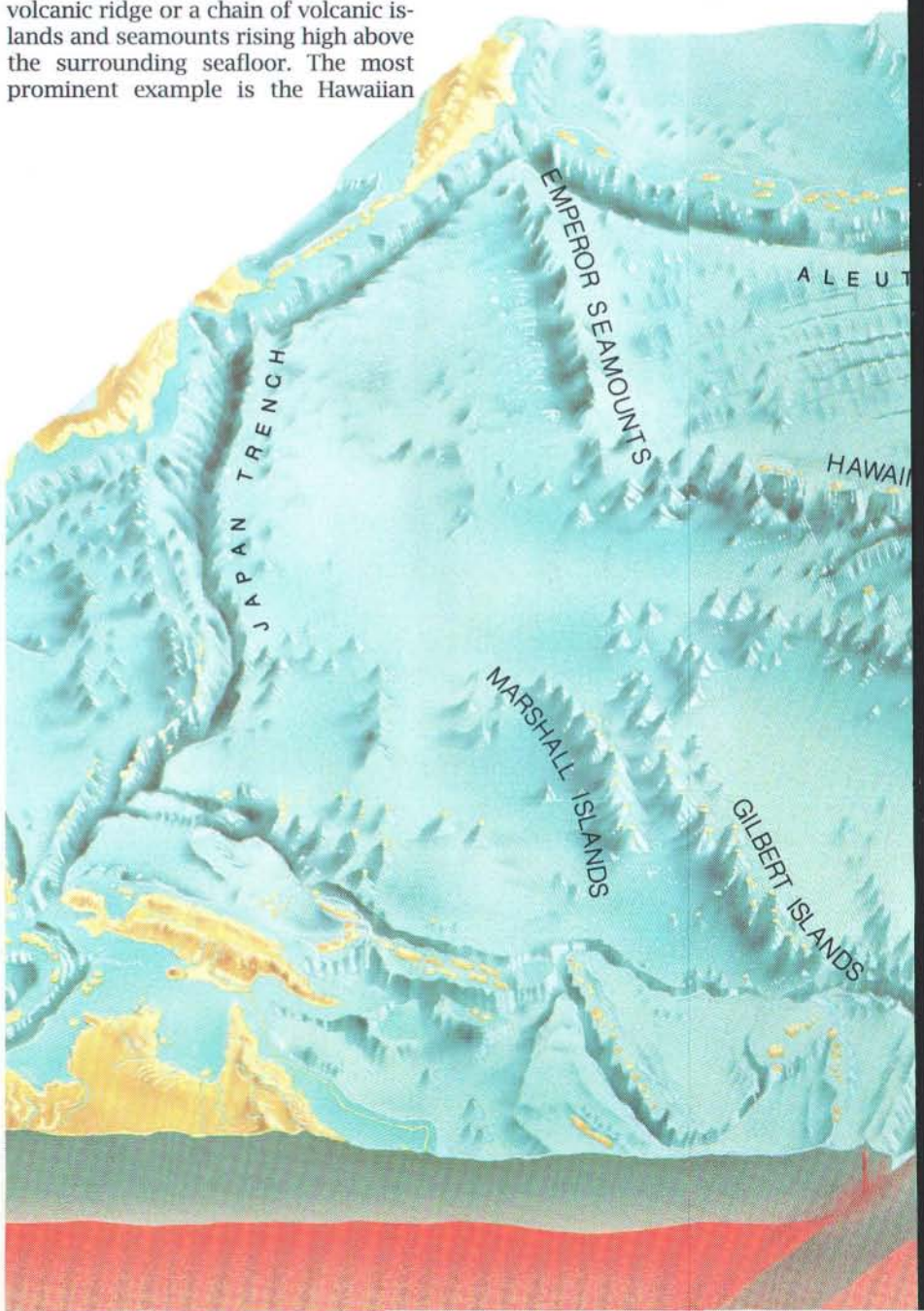
The Earth's Hot Spots

These plumes of hot rock welling up from deep in the mantle are a key link in the plate tectonic cycle. The marks they leave on passing plates include volcanoes, swells and mid-ocean plateaus

by Gregory E. Vink, W. Jason Morgan and Peter R. Vogt

From deep inside the earth's mantle, isolated, slender columns of hot rock rise slowly toward the surface, lifting the crust and forming volcanoes. The plumes well up all over the world, under oceans and continents, both in the center of the mobile plates that make up the earth's outer shell and at the mid-ocean ridges where two plates spread apart. The marks they leave at the surface are superposed on the grand effects of plate motion. Volcanic eruptions and earthquakes associated with plumes occur far from plate boundaries, the site of most such activity; the upwelling currents also form broad anomalous swells in the ocean floor and in continental terrain. These isolated areas of geologic activity are called hot spots.

Mantle plumes are relatively stationary, and so the crustal plates drift over them. Often the passage of a plate over a hot spot results in a trail of identifiable surface features whose linear trend reveals the direction in which the plate is moving. If the plate is oceanic, the hot-spot track may be a continuous volcanic ridge or a chain of volcanic islands and seamounts rising high above the surrounding seafloor. The most prominent example is the Hawaiian



GREGORY E. VINK, W. JASON MORGAN and PETER R. VOGT specialize in marine geophysics. Vink is director of planning at the Incorporated Research Institutes for Seismology and visiting lecturer at Princeton University. Morgan is professor of geophysics at Princeton University, and Vogt is a staff geophysicist at the Naval Research Laboratory. Vink got his B.A. at Colgate University in 1979 and his Ph.D. from Princeton in 1983; his dissertation described the tectonic evolution of the Norwegian-Greenland Sea and the Arctic Ocean. Morgan went to the Georgia Institute of Technology as an undergraduate and received a Ph.D. in physics from Princeton in 1964. Vogt was educated at the California Institute of Technology and the University of Wisconsin at Madison, which granted him a Ph.D. in oceanography in 1968. He worked in the U.S. Naval Oceanographic Office until 1975 and then joined the Naval Research Laboratory; he holds a concurrent post at the University of Oslo. In 1975 Vogt served as co-chief scientist on the *Glomar Challenger* deep-sea drilling expedition to the North Atlantic.

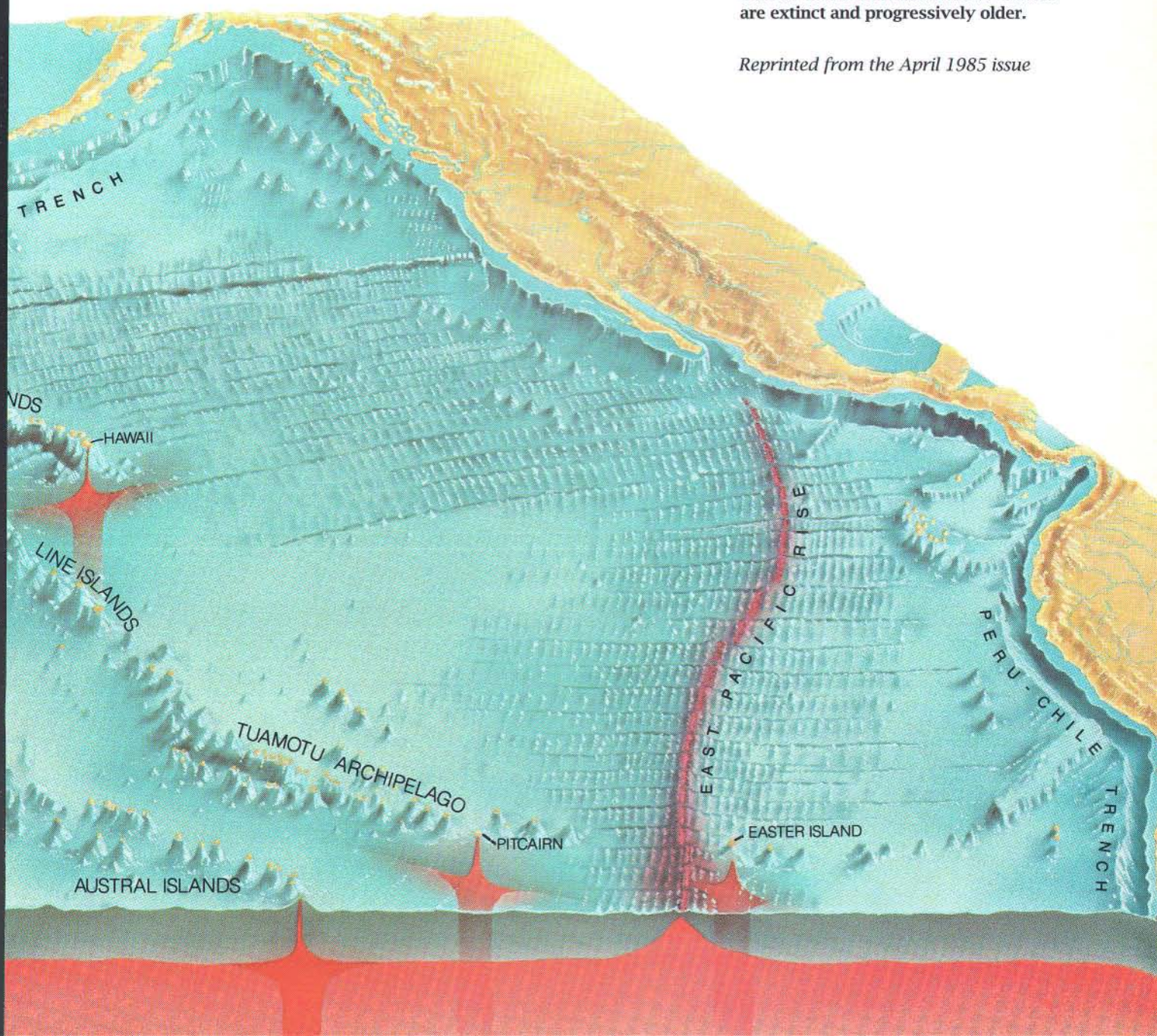
Islands; it was a visit there that led J. Tuzo Wilson of the University of Toronto to put forward the concept of hot spots in 1963.

Wilson noticed that to the west of Hawaii the islands disappear into atolls and shoals, indicating they are progressively more eroded and therefore older. The same observation had been made more than a century earlier by the American geologist James Dwight Dana, but Wilson was the first to interpret the age progression as evidence of continental drift. He proposed that the island chain had been formed by the westward motion of a crustal slab over "a jetstream of lava" now situated under Hawaii itself, at the eastern end of the chain. The proposal came at a time when textbooks, including one co-authored just three years earlier by Wilson himself, mentioned continental drift only as an intriguing idea that had been advanced in the 1920s but was later discredited.

In the past two decades the idea has become generally accepted as part of the theory of plate tectonics. The earth's crust is now known to be embedded in the rigid plates of the lithosphere, which is between 100 and 150 kilometers thick under continents and about half as thick under oceans; the continual motion of the plates over the

MOTION OF THE PACIFIC PLATE over three fixed mantle plumes has produced three parallel island chains: the Hawaiian Islands and Emperor Seamounts, the Tuamotu and Line islands, and the Austral, Gilbert and Marshall islands. The chains lie in the center of the plate, proving they were formed by a mechanism different from the one that built the volcanic island arcs of the western Pacific, which are associated with the subduction of the plate at oceanic trenches. The plumes originate deep in the mantle, and their surface tracks reveal the path of the slowly moving plates. About 40 million years ago the Pacific plate switched to its present westward course from a more northerly heading; the change shows up as a bend in the hot-spot chains. Active volcanoes, such as Kilauea on Hawaii, are at the southeastern end of the chains. To the northwest the volcanoes are extinct and progressively older.

Reprinted from the April 1985 issue



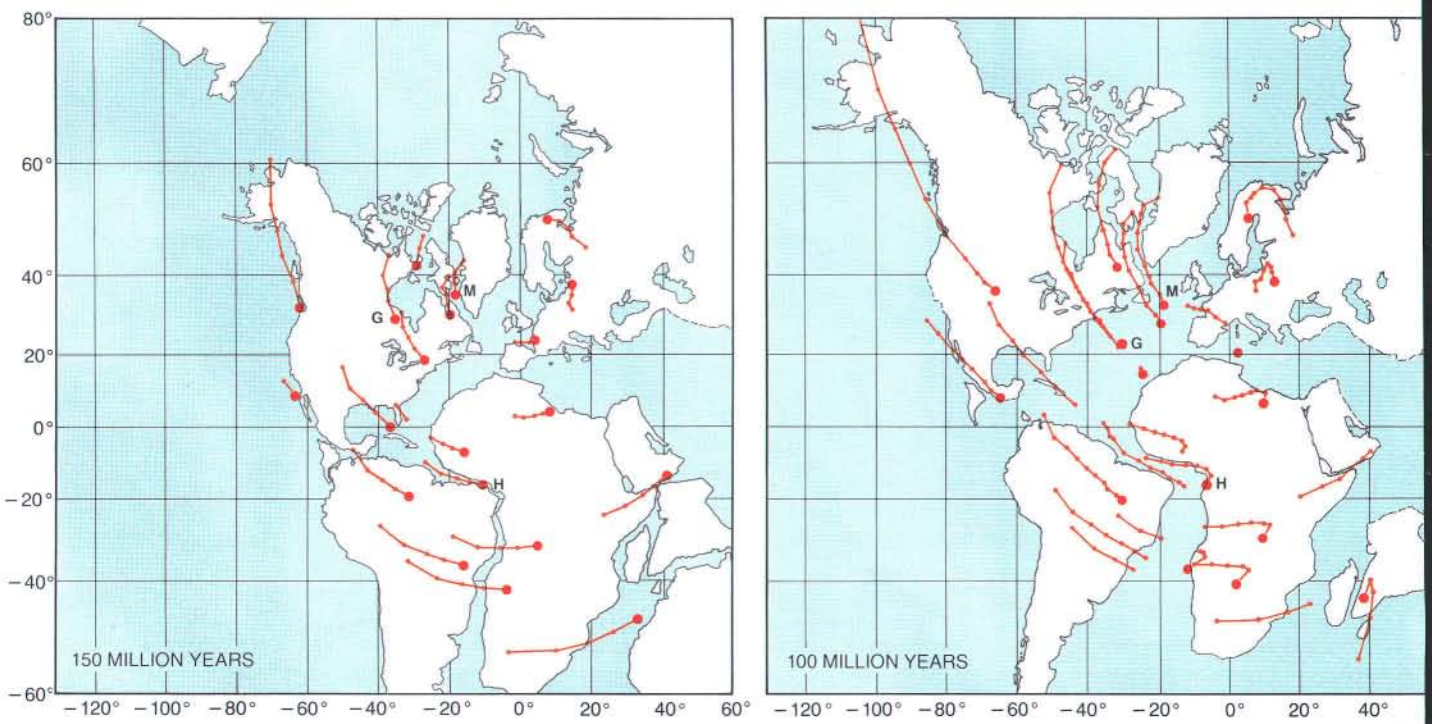
partially molten asthenosphere (the portion of the mantle extending to a depth of roughly 200 kilometers) explains the development of ocean basins and the formation of mountain ranges. A major task of contemporary geophysics is to understand how these surface processes are related to the slow convective "creep" of hot rock in the underlying mantle. Hot spots are an important part of this connection.

Indeed, if the upwelling plumes were to stop, the plates would grind to a halt. Ultimately the energy that drives plate motion is the heat released by the decay of radioactive elements deep in the mantle. The plumes provide an efficient way of channeling the heat toward the surface. Their efficiency is attributable to a property of mantle rock: its viscosity, or resistance to flow, is reduced dramatically by relatively small increases in temperature (say, 100 degrees Celsius) or in the content of volatile elements such as water. Less viscous material produced by variations in temperature or volatile content tends to collect and rise toward the surface through a few narrow conduits, much as oil in an underground reservoir rises through a few boreholes.

It would be misleading, however, to say that the plumes propel the plates. Rather the two are different parts of the same convective cycle. As plates spread apart at a mid-ocean ridge, molten rock from the asthenosphere wells up at the spreading axis to form oceanic crust; the new lithosphere cools as it moves away from the ridge and is eventually destroyed at oceanic trenches, where two plates collide and one of them sinks deep into the mantle. The deep mantle feeds the plumes. The plumes in turn empty matter heated by radioactivity into the asthenosphere, which in addition to serving as the source of new seafloor provides a hot and fluid layer for the plates to glide across. The asthenosphere is constantly being destroyed as it cools and attaches to the base of the lithosphere; the boundary between the two layers is essentially a thermal one. Were it not replenished by the plumes, the asthenosphere would soon vanish, and the motion of the plates would stop.

It is worth emphasizing that this "plume model" of the convective circulation in the mantle is just that: a model. The plumes have not been observed directly. The deep mantle can be explored only through the analysis of earthquake waves, and so far the resolution of seismic studies has not been good enough to detect plumes; the upwelling currents may be just a few hundred kilometers in diameter and only moderately different from their surroundings in temperature and density (the properties that determine the seismic-wave velocity in a region).

The indirect evidence for deep-mantle plumes, however, is substantial. Satellite measurements of the earth's gravity field have shown hot spots to be areas of anomalously high gravity and thus of excess mass; the excess mass can be attributed to broad bulges in the surface produced by the upwelling plumes. A second line of evidence comes from geochemical studies of basalts erupted at hot-spot volcanoes. Compared with the basalts dredged from mid-ocean ridges, these rocks are enriched in volatile elements



HOT-SPOT TRACKS reveal how the plates have moved with respect to the earth's interior during the opening of the Atlantic Ocean. Because spots, represented in this illustration by large dots, are anchored deep in the mantle, they remain relatively fixed, that is, their latitude and longitude remain unchanged. The tracks consist of extinct volcanoes, magma intrusions and swells in the crust formed by the upwelling plumes and then carried away by the plates. Each small dot represents 10 million years of plate motion. In reconstructing the plate motions, one begins with one or two well-defined tracks, such as that of the Great Meteor hot spot (*G*), which also formed the New England Seamounts and magma intrusions in the White Mountains. The tracks of other hot spots are then calculated from the reconstructions, which must fit

and in other elements such as potassium that are "incompatible" with the crystals of ordinary mantle rock. They also contain anomalous amounts of isotopes derived from radioactive decay processes. The differences in composition suggest that hot-spot lavas are derived from rock welling up from below the asthenosphere, which feeds the oceanic spreading centers. According to the plume model, as material from the deep mantle flows into the asthenosphere, the part rich in volatiles and other incompatible elements melts, and some of it rises to the surface at hot-spot volcanoes.

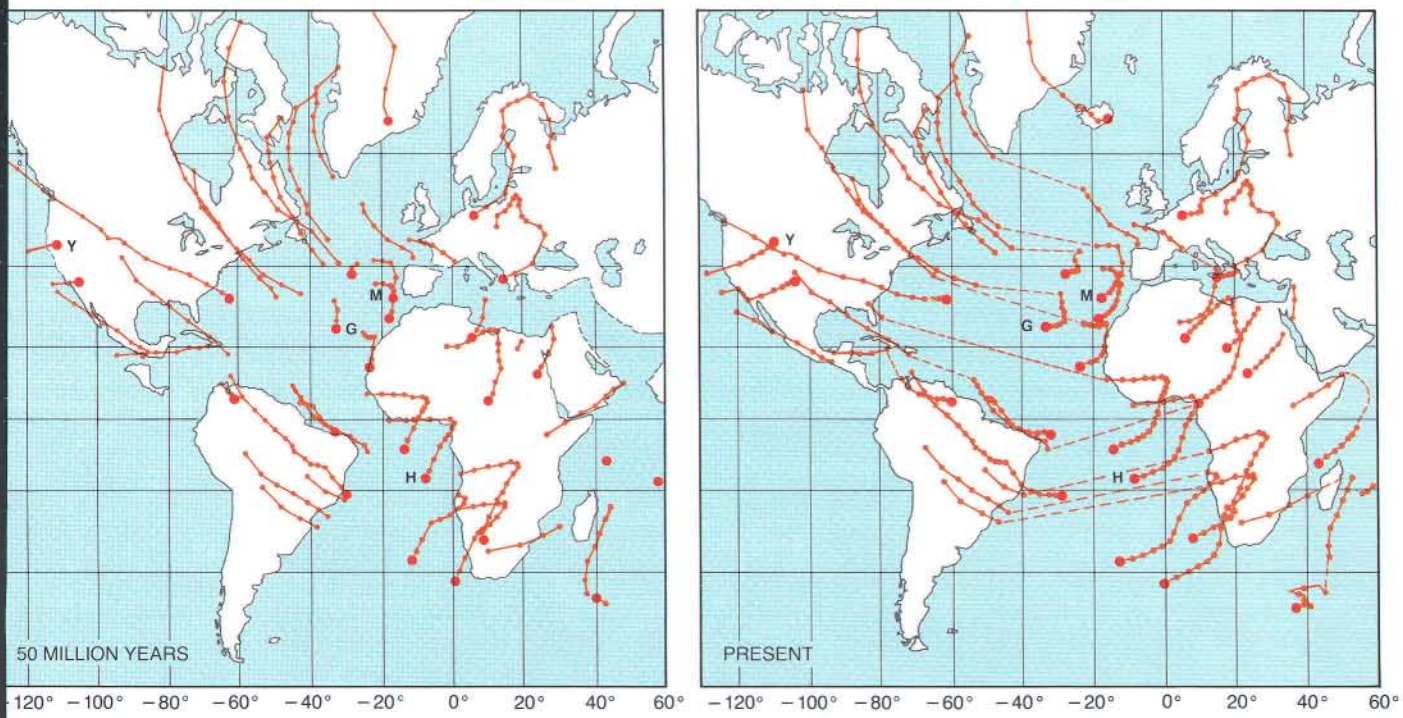
Recent advances in seismology encourage the hope that someday workers will observe the plumes directly [see "Seismic Tomography," by Don L. Anderson and Adam M. Dziewonski; SCIENTIFIC AMERICAN, October 1984]. In particular, a proposed new global network of seismometers may improve the resolution of seismic studies to the point where it is possible to determine the size of plumes and the depth of their roots.

The plumes are certainly not uniform; differences in their isotope signatures imply that they come from various depths. Comparisons of the volume and frequency of eruptions at different hot spots indicate they also come in a range of sizes. Furthermore, individual plumes are not immutable. After examining the volume of rock extruded along the Hawaiian hot-spot track, one of us (Vogt) has suggested that the discharge rate of a plume may vary over time. Geochemical evidence supports the conclusion. Jean-Guy E. Schilling of the University of Rhode Island has proposed that plumes consist of rock rising in blobs rather than in a continuous flow.

Sometimes a hot spot may fade away entirely, and new ones may be formed; from the tracks, it appears the typical life span of a plume is on the order of 100 million years. Moreover, the position of a hot spot seems to change slightly. As a result, the tracks on the surface are not all as neatly linear as the Hawaiian chain.

Compared with the plates, however, the mantle plumes are relatively stationary. The first evidence of their fixity came in 1970. One of us (Morgan) showed that three volcanic island groups in the Pacific—the Hawaiian Island-Emperor Seamount chain, the Tuamotu Archipelago-Line Island chain and the chain formed by the Austral, Gilbert and Marshall islands—are approximately parallel and could all have been formed by the same motion of the Pacific plate over three fixed hot spots. In each case, the most recent volcanic activity has taken place near the southeastern end of the chain, and the islands and seamounts get progressively older to the northwest. The Pacific plate is currently moving toward the northwest; it switched to that course from a more northerly heading about 40 million years ago. The course change shows up as a bend in the hot-spot tracks.

Because the motion of the hot spots is insignificant, they provide a worldwide reference frame for tracing the absolute motions of the plates with respect to the earth's interior. For some time, workers have mapped the paths of the plates in relation to one another



the relative plate motions derived from the historical evidence of seafloor spreading. When the mid-ocean ridge separating two plates drifts over a plume, the track continues on the other plate but is interrupted (*broken lines*) by seafloor formed at the ridge after it passed over the hot spot. A plate motion is a rotation, and so the tracks approximate concentric circles rather than parallel straight lines. Along the Madeira (*M*) and St. Helena (*H*) tracks, continents have later rifted apart; the plumes may promote rifting by thinning a passing plate. The Snake River plain, where the lithosphere has been weakened by the track of the Yellowstone hot spot (*Y*), may be the site of a future rift. Not all hot spots are present in each reconstruction because over the millennia new ones form and old ones fade away.

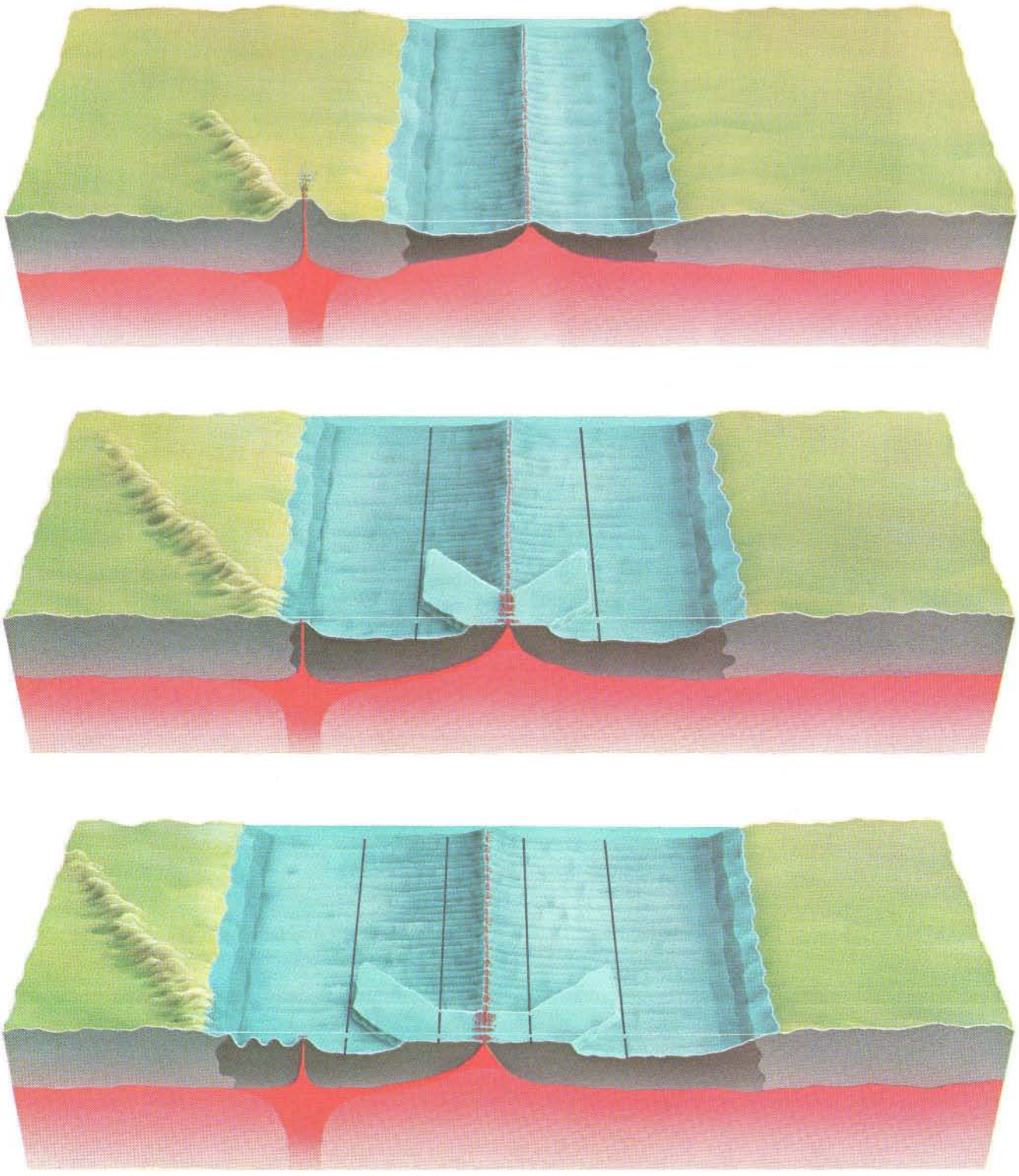
MID-ATLANTIC RIDGE is positioned over several hot spots; the flow from these plumes adds to the normal upwelling of magma at the Ridge, producing thicker crust. In the computer-plotted topographical map, brown regions are shallow and green regions are deep. Iceland is perched on the ridge axis and also has a large hot spot under its southeast coast; the plume has raised the crust above sea level by lifting and thickening it. The tapered structure of the ridge segment south of Iceland, called the Reykjanes Ridge, reflects the flow of plume material down the axis. Similar topography southwest of the Azores suggests material from that hot spot is also flowing along the Ridge. The Iceland hot spot may have formed the plateau southeast of Iceland (including the Faeroe Islands) by feeding a now extinct spreading axis at the center of the plateau. William F. Haxby of Columbia University's Lamont-Doherty Geological Observatory prepared the map from data compiled by Joseph E. Gilg and Roger Van Wyckhouse of the U.S. Naval Oceanographic Office.

and have thereby been able to reconstruct the opening of ocean basins. The boundaries between plates—the ridges and trenches—also move, however, and so the relative motions do not reveal where on the globe a plate was at a given time. Nor do they indicate whether two diverging plates have been moving at the same speed or whether instead one plate has remained stationary. Such questions can be answered by converting the known relative motions into absolute motions in the hot-spot reference frame, in which each hot spot occupies an unchanging latitude and longitude.

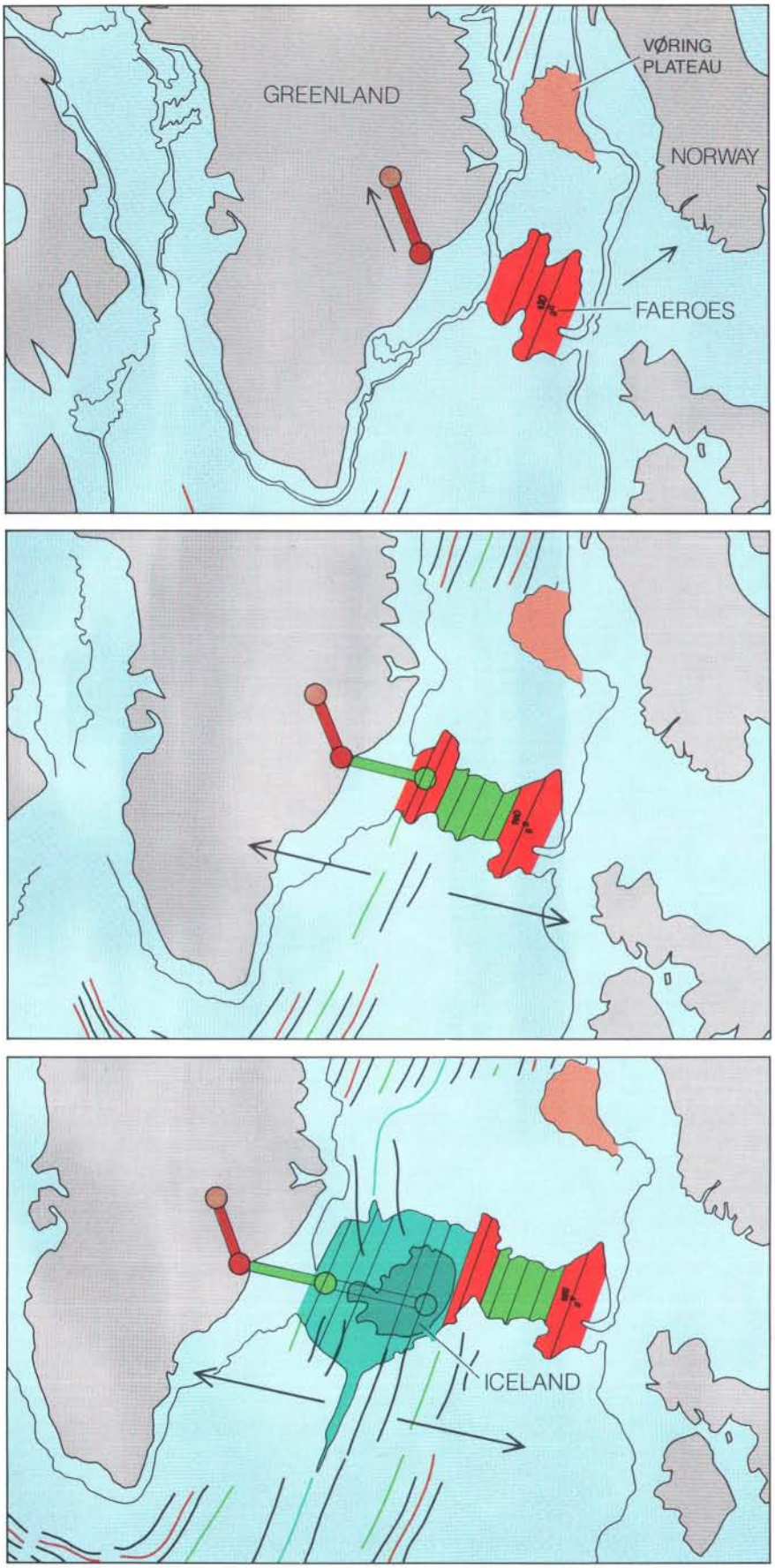
The relative motion of diverging plates—the seafloor-spreading history—is determined through the analysis of magnetic anomalies in the seafloor. Throughout geologic history, at regular intervals averaging about 100,000 years, the earth's magnetic field has reversed its polarity, for reasons that are poorly understood. A record of these reversals is preserved in the oceanic crust. The magnetic minerals in lava erupting from mid-ocean ridges align themselves with the prevailing field, and as the molten rock cools and solidifies, the field direction is permanently locked in the crust.

The magnetized crust is transported away by the diverging plates in bands that roughly parallel the ridge axis. Each band has a characteristic magnetic anomaly and is made up of crust formed at the same time, and so the bands are called magnetic isochrons. The age of various isochrons, and therefore the seafloor-spreading rate, has been established through radiometric dating of rocks retrieved in deep-sea drilling expeditions. By superposing corresponding isochrons from opposite sides of the spreading axis, one can reconstruct the relative position of the plates at the time the isochron pair was formed. (The superposition in effect removes from the map all seafloor created after the particular magnetic reversal.)

If the motion of one of the plates over the plumes is known, then their relative motion allows the path of other plates in the hot-spot reference frame to be deduced. The customary procedure is to begin with a well-defined hot-spot track on one plate—say, a chain of seamounts—and then adjust the more ambiguous tracks until the "best fit" is achieved: the absolute plate motions that best satisfy the constraints established by the hot-



HOT SPOT MAY FEED A RIDGE from a distance, thickening the crust and forming an oceanic plateau. Early in the opening of the ocean basin (*top*), the hot spot is under a thick continental plate moving to the northwest; material from the plume cannot yet reach the spreading center. Millions of years later (*middle*) the motion of the plates has brought the ridge closer and has carried continental shelf over the hot spot. Material from the plume has begun to flow along the lithosphere to the nearest section of the rise. As the excess material erupts, it is carried away on the plates; the V shape of the resulting plateau reflects both the spreading of the plates away from the ridge and their motion with respect to the hot spot. A change in plate motion (*bottom*) forms a bend in the hot-spot track and in the plateau.



spot evidence and the relative motions.

Using this procedure, we have reconstructed the opening of the Atlantic and Indian oceans. The reconstructions can be tested: surface features along the hot-spot tracks must by their nature and age fit the hypothesis that they were formed by the passage of a plate over an upwelling plume. This should be true not only along the well-defined parts of the tracks but also in regions where the tracks have simply been extrapolated from the calculated plate motions and evidence of hot-spot activity has not previously been observed.

Although the available data are fragmentary (particularly concerning the ages of seafloor features), in general the reconstructions pass the test. A good example is the track of the hot spot that formed the Great Meteor Seamount south of the Azores [*see illustration on pages 134 and 135*]. Two hundred million years ago the area northwest of Hudson Bay on the Arctic Circle was over the Great Meteor plume; 50 million years later the hot spot was under Ontario. The exposure of the Canadian shield from Manitoba to Ontario can be attributed to uplifting of the crust by the plume: in an uplifted area, sediment covering the basement rocks is more likely to be eroded away over time.

One hundred million years ago the track had reached the young and narrow Atlantic off Cape Cod. The passage of New Hampshire over the hot spot is recorded by magma intrusions in the metamorphic rock of the White Mountains; the intrusions are between 100 and 124 million years old. For the peri-

GREENLAND-FAEROE PLATEAU might have been formed by the Iceland hot spot. The parallel colored lines are magnetic isochrons used to reconstruct past positions of the plates. Fifty million years ago (*top*) the hot spot was under the coast of Greenland and began feeding the ridge. The V shape of the plateau reflects the hot-spot track. By 36 million years ago (*middle*) the plates had changed course, as reflected in the new section of the plateau. At about that time, the spreading axis moved west over the hot spot, which by then was under oceanic lithosphere. Hot spot–fed spreading has continued in the west until now (*bottom*), forming Iceland. At an earlier time, when it was close to a northern ridge, the plume may have built the Vøring Plateau.

od from about 100 to 80 million years ago, the track follows the trend of the New England Seamounts. Based on radiometric dating of rocks collected from the seamounts, Robert A. Duncan of Oregon State University has shown that the volcanoes get progressively younger toward the southeast along the chain. Their ages coincide with their passage over the hot spot. From the ages and the distances between the seamounts, Duncan has calculated the velocity of the North American plate during that period: about 4.7 centimeters per year.

Approximately 80 million years ago the Mid-Atlantic Ridge migrated westward over the plume. The track continues on the African plate and ends at the Great Meteor Seamount. At present the hot spot should be about 500 kilometers southwest of Great Meteor. Although there is a swell in that region of the seafloor, there is no sign of current volcanism; the plume may have become inactive.

A swell in the ocean floor, like an exposed continental shield, is an area of uplifted crust. Some time ago Robert S. Detrick and S. T. Crough, then at the University of Rhode Island, proposed that a plume produces uplift not by bending the lithosphere but by thinning it, replacing cold, dense lithosphere with hot, buoyant rock from the asthenosphere. After passing over an active hot spot, both seafloor and continental swells presumably cool and gradually sink back to their former altitude. Swells on the seafloor are interruptions of the process in which the lithosphere cools, thickens and sinks as it moves away from a mid-ocean ridge, eventually plunging into the asthenosphere at a trench.

The hot-spot anomalies, however, are by no means insignificant interruptions. There are some 40 active hot spots, and the swells associated with them have an average diameter of about 1,200 kilometers. Thus swells cover roughly 10 percent of the earth's surface. This observation led Crough and Richard Heestand of Princeton University to suggest that the depth of the seafloor in a particular region is controlled not only by the progressive cooling of the lithosphere but also by the time elapsed since the region passed over a hot spot.

In the same way, hot spots could control the thickness of the continental lithosphere. Moreover, the thinning and weakening of continental plates by mantle plumes may produce more dramatic effects than the exposure of basement rock: it may cause them to rift apart. In the early 1970s Kevin C. Burke of the State University of New York at Albany noticed that some hot spots are associated with three-arm rift systems, in which two of the arms have formed a plate boundary, whereas the third has failed. The failed rifts form valleys extending into the continents; an example is the Niger River Valley.

The reconstructions of the Atlantic opening reveal a number of hot-spot tracks along which continents have subsequently broken up, probably millions of years after the plates passed over the plumes. The track of the hot spot that formed the Madeira Islands, for example, runs between the west coast of Greenland and the east coast of Baffin Island and Labrador; the plume that created St. Helena can be traced along the south coast of West Africa and the north coast of Brazil. In the future, a rift may develop in the Snake River plain, where the North American plate has been weakened by the track of the hot spot now under Yellowstone National Park.

Mantle plumes explain much of the geologic activity in the center of the plates. As the plates move over the hot spots, however, so do the plate boundaries, including the mid-ocean ridges; unlike the hot spots, the ridges are not anchored deep in the mantle. What happens when a plume is under or near a spreading axis?

A plume directly under a spreading center augments the flow of molten rock welling up from the asthenosphere to form new crust. The crust over the hot spot is therefore thicker than it is along the rest of the ridge, and the result is a plateau rising above the surrounding seafloor. The most striking example is Iceland, a hot-spot island that straddles the Mid-Atlantic Ridge: there the upwelling is so intense and the crust so exceptionally thick that the plateau is above sea level. Geochemically the Icelandic crust is distinctly different from typical oceanic crust; it shows clear evidence of a hot-spot contribution. Gravity measurements indicate that the core of the plume is under the southeastern part of the island. The volcanic peaks there are visible signs of a powerful upwelling current: as much as 5,500 feet high, they are covered by the Vatnajökull glacier. (In 1918 an eruption under the glacier unleashed a flood of meltwater at a discharge rate 20 times that of the Amazon River.)

Some of the material in the strong Iceland plume also seems to spread out under the lithosphere. The lithosphere slopes upward toward a spreading axis, and one of us (Vogt) has proposed that the axis north and south of Iceland has acted as a pipeline, channeling partially molten rock away from the hot spot. In both directions along the ridge the excess plume material produces abnormally elevated topography out to a distance of approximately 1,500 kilometers. To the south of Iceland the broad plateau tapers to form the typical Mid-Atlantic Ridge. The tapered structure probably arises from the fact that most of the volatile-rich, easily melted plume rock is used up near Iceland. Indeed, Schilling has found that the chemical composition of basalts dredged from the Ridge becomes progressively more like "normal" oceanic crust with increasing distance from Iceland, suggesting that the relative contribution of the hot spot gradually declines.

On the flanks of the ridge south of Iceland there are symmetric pairs of secondary ridges. Each pair forms a southward-pointed V whose apex is on the spreading axis. These features could have been produced by "waves" of intensified flux or of unusually hot and buoyant material from the plume. A wave traveling down the ridge would generate anomalously thick crust, affecting the area nearest the hot spot first. The elevated crust would then be carried away on each side of the axis by the spreading plates, forming the V-shaped secondary ridges. From the known spreading rate and the angle between the secondary ridges and the spreading axis, one can estimate the speed of the plume material; it seems to flow down the axis at a rate of five to 20 centimeters per year.

Because the mid-ocean ridges move, a hot spot is unlikely to be situated under a spreading center for more than a geologically brief period. It is conceivable, however, that a plume might feed a spreading axis from a distance, provided it is close enough to the region in which the base of the lithosphere slopes up toward the axis. This concept helps to explain certain unusual surface features in the Iceland area.

The plateau that includes Iceland stretches from Greenland in the west

to the Faeroe Islands in the east. The section of the plateau east of Iceland and east of the current spreading center has long puzzled geologists. Its linear trend suggests a hot-spot origin. Yet it could not simply have been formed by the motion of a plate over a fixed plume, because it does not coincide with the track of the Iceland hot spot, which is known from the reconstructions of the early Atlantic. Some workers have interpreted this as a sign that the hot spot has not remained stationary but has instead wandered about, forming the plateau by occasionally punching through the plate. The argument implies that the reconstructions are inaccurate: if plumes are not fixed, they provide no absolute reference frame for mapping plate motions over the mantle.

Our own hypothesis is that the Iceland hot spot has remained stationary and that the Iceland-Faeroe plateau section was made by rock flowing eastward from the hot spot to a now extinct spreading center. The hypothesis can be tested. Presumably the plume would feed the closest point on the ridge. Thus, at any time during the formation of the plateau, a line representing the shortest distance from the plume to the ridge should intersect the center of the plateau. The plateau would be symmetric about the ridge axis but not necessarily perpendicular to it. With respect to the hot spot, the plates might have a component of motion parallel to the axis, and the orientation of the plateau would be obtained by adding that component to the relative motion of the plates (perpendicular to the axis). Finally, the age of the plateau at any point would be the same as that of the surrounding seafloor, because the two were formed at the same time. None of these predictions would hold if the plateau were formed by a wandering hot spot that was not feeding a ridge.

To test the model, one of us (Vink) reconstructed the opening of the Norwegian-Greenland Sea and the formation of the plateau. The method is the same as that used for reconstructing the early Atlantic: superposing magnetic isochrons reveals the relative position of the plates at the time of a given magnetic anomaly, and the hot-spot track shows the plate motions in the hot-spot reference frame.

During the early opening of the basin, some 50 to 60 million years ago, the Iceland hot spot was under eastern Greenland. Its southerly track reflects the northward motion of the Greenland plate. The passage of the plate over the plume probably produced the extensive igneous rock formations southwest of Scoresby Sound, which from radiometric evidence are judged to be roughly 55 million years old. About 50 million years ago the Greenland continental shelf moved over the hot spot. Excess plume material could have begun flowing along the base of the oceanic lithosphere to the spreading center, and the plateau would have started to form. The Faeroe Islands, now at the eastern end of the plateau, would have been created first; their basalts are between 50 and 60 million years old. In the reconstruction of the period the nascent plateau is roughly symmetric about the spreading axis, and the V shape of its northern edge reflects the northerly motion of the plates with respect to the hot spot.

By 36 million years ago the plates had switched to a more westerly course, causing the hot-spot track to bend to the east. The change is apparent in the geometry of the plateau: the V is split by a younger segment with an east-west heading, perpendicular to the spreading axis. The plateau remains symmetric about the axis, and a line from the hot-spot position intersects the axis at the center of the plateau. Both observations indicate the plume was continuing to channel molten rock to the ridge.

By that time the hot spot was under oceanic lithosphere, which is somewhat thinner than continental lithosphere. The plume would have thinned it further. Our model assumes that the ridge subsequently jumped to the area of weakened lithosphere, leaving an extinct spreading center on the eastern section of the plateau. Although the existence of such a relic is still being debated, geologic activity seems indeed to have ceased in the east at about the time the spreading axis would have jumped to the west; rocks collected from a drill hole near the center of the eastern section are roughly 40 to 43 million years old.

Seafloor spreading continued at the western end of the plateau. With the hot spot positioned under the spreading axis, plume material began to flow down the axis, giving the ridge its present tapered structure to the south. The westward-moving plates soon pushed the axis off the hot spot, but the plume continued to feed the ridge. The oldest outcrops on Iceland are found near the east and west coasts; as one would expect on an island formed at a spreading axis, their ages suggest the island was born between 16 and 12 million years ago. Iceland remains geologically active. In the past few million years eastward movements of the spreading axis have once again placed the ridge over the hot spot.

The reconstructions show that a fixed Iceland hot spot could well have produced the observed geometry of the Greenland-Faeroe Plateau. It may also have formed the Vøring Plateau, even though that is now 500 kilometers north of Iceland. The hypothesis rests on the assumption that a plume will always feed the closest section of a spreading axis. Just before the formation of the Greenland-Faeroe Plateau, when the hot spot was still under Greenland, it may have been closest to a northern ridge segment. During that period, it could have produced the Vøring Plateau. The northerly motion of the Greenland plate later brought the southern spreading axis closer to the hot spot, and so the plume switched targets.

Like plate tectonics itself, the notion of hot spots is a simple but powerful concept. It explains many features of the earth's surface that once seemed disparate, and further research will undoubtedly lead to the attribution of other effects to upwelling plumes in the mantle. At the same time, the concept is appealingly intuitive. Indeed, it is only embellishing the truth a little to suggest the Hawaiians recognized the track of their hot spot centuries before it caught the attention of modern geologists. According to Hawaiian legend, Pele, the fiery-eyed goddess of volcanoes, originally lived on Kauai, at the western end of the island chain. When the god of the sea evicted her, she fled to Oahu. Forced again to flee, she continued to move east, to Maui, and finally to the island of Hawaii. She now seethes in the crater at Kilauea.

FURTHER READING

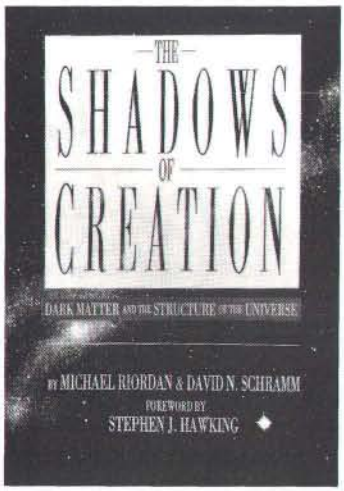
THE ICELAND MANTLE PLUME: STATUS OF THE HYPOTHESIS AFTER A DECADE OF NEW WORK. Peter R. Vogt in *Structure and Development of the Greenland-Scotland Ridge.* Edited by Martin H. Bott and Svend Saxov. Plenum Publishing, 1983.

HOTSPOT TRACKS AND THE EARLY RIFTING OF THE ATLANTIC. W. Jason Morgan in *Tectonophysics,* Vol. 94, No. 1–4, pages 123–139; May 1, 1983.

A HOTSPOT MODEL FOR ICELAND AND THE VØRING PLATEAU. Gregory E. Vink in *Journal of Geophysical Research,* Vol. 89, No. B12, pages 9949–9959; November 10, 1984.

W. H. Freeman and Company

CHALLENGING THE FRONTIERS OF SCIENCE



THE SHADOWS OF CREATION

Dark Matter and the Structure of the Universe

Michael Riordan and David N. Schramm

Foreword by Stephen J. Hawking

Join Michael Riordan and David Schramm as they explore the biggest scientific mystery of the decade: dark matter—the more than ninety percent of the universe that is unseen and unknown. What is this dark matter? Why do cosmologists say this undetected "missing mass" must exist? What does it tell us about our universe? Riordan and Schramm explore these questions as they describe efforts underway at observatories, laboratories, and particle colliders to determine the nature of dark matter.

1991, approx. 224 pages, 63 illus.; Cloth: 2157-0 $18.95 *SPECIAL PRICE* $14.21

FRACTALS, CHAOS, POWER LAWS

Minutes from an Infinite Paradise

Manfred Schroeder

The distinguished physicist Manfred Schroeder traces the development of self-similarity in mathematics and provides keen insights into the analytical power with which it describes seemingly different phenomena. *Fractals, Chaos, Power Laws* reveals the fascinating implications of self-similarity and explores the surprising presence of self-similar symmetries in physics, chemistry, music, and the visual arts. The text is complemented by over 200 illustrations, including 8 pages in full color.

1991, 432 pages, 246 illus., 8 page color insert
Cloth: 2136-8 $32.95 *SPECIAL PRICE* $24.71

HOW COMPUTERS PLAY CHESS

David Levy • Monroe Newborn

Computer chess is more intriguing than ever. Here, Levy and Newborn explore the world of computer chess, from the history to how the computer decides where to move. The book examines most aspects of programming computers to play chess. *How Computers Play Chess* is the definitive book for the chess aficionado and computer buff alike.

Computer Science Press
1991, 264 pages, 160 illustrations
Cloth: 8239-1 $23.95 *SPECIAL PRICE* $17.99
Paper: 8121-2 $11.95 *SPECIAL PRICE* $8.99

VISUALIZATION

The Second Computer Revolution

Richard Mark Friedhoff

Computer graphics are revolutionizing every field that utilizes images. *Visualization* is the first book to describe in nontechnical language the technologies underlying the digital processing, generation, and manipulation of images. An indispensable resource for learning, *Visualization* also reveals how the computer is making possible new thinking in the arts, architecture, science, medicine, engineering, communications, and entertainment.

1991, 216 pages, 200 illus., 136 in full color
Paper: 2231-3 $25.95 *SPECIAL PRICE* $19.46

AARON'S CODE

Meta-Art, Artificial Intelligence, and the Work of Harold Cohen

Pamela McCorduck

Aaron's Code tells the story of Aaron, a computer program created by artist Harold Cohen. The first profound connection between art and computer technology, Aaron is programmed to make creative drawings without human intervention. Pamela McCorduck explores Cohen's pioneering work in seizing the emblematic machine of the 20th century and using it to make art history.

1990, 225 pages, 77 illus. plus color inserts
Cloth: 2173-2 $24.95 *SPECIAL PRICE* $18.71

Your satisfaction matters. Return any book in good condition within 15 days of receipt and we'll send you a prompt unconditional refund.

THE HOUR OF OUR DELIGHT

Cosmic Evolution, Order, and Complexity

Hubert Reeves

One of France's most distinguished researchers and a well-known popularizer of science examines the evolution of our universe from scientific and philosophical perspectives. Our complexity has given us all the remarkable products of modern technology—*and* the means to destroy ourselves. Will human intelligence continue its evolution or descend into self-destruction? This thoughtful analysis was a bestseller in France, where it won the prestigious Prix Blaise Pascal.

1991, approx. 256 pages, 21 illustrations
Cloth: 2220-8 $17.95 *SPECIAL PRICE* $13.46

A video introduction with Mandelbrot and Lorenz

FRACTALS

An Animated Discussion with Edward Lorenz and Benoit B. Mandelbrot

H.-O. Peitgen • H. Jürgens • D. Saupe • C. Zahlten

A fascinating visual exploration, this new 63-minute video is an insightful combination of full-color animated sequences and intriguing interviews. The film turns the Mandelbrot set and the Lorenz attractor into visible objects as their discoverers, Benoit B. Mandelbrot and Edward Lorenz, discuss details of their work. It also features new computer-graphic illustrations of chaos and self-similarity, as well as music composed according to fractal principles.

1990, 63-minute VHS video 2213-5 $59.95 *SPECIAL PRICE* $47.96

MIND SIGHTS

Original Visual Illusions, Ambiguities, and Other Anomalies, with a Commentary on the Play of Mind in Perception and Art

Roger N. Shepard

Part autobiography, part artist's portfolio, and part essay on perception and the psychology of art, *Mind Sights* proves both stimulating to the eye and provocative to the mind. Renowned psychologist Roger N. Shepard introduces his wonderfully original drawings of visual tricks (many never before published), discusses the origin of his scientific and artistic work, and shares his reflections on the nature of art, perception, and the mind.

1990, 228 pages, 200 illustrations
Cloth: 2134-1 $24.95 *SPECIAL PRICE* $18.71
Paper: 2133-3 $14.95 *SPECIAL PRICE* $11.21

Now in paperback!

KEEP YOUR EYE ON THE BALL

The Science and Folklore of Baseball

Robert G. Watts • A. Terry Bahill

Could Sandy Koufax's curve really "fall off a table?" Why does a well-pitched knuckle ball silence so many great bats? Were the hitters of yesterday really better than those of today? Engineers Watts and Bahill put some of baseball's cherished myths to the test of scientific scrutiny in a highly informative and entertaining guide.

1990, 212 pages, 101 illustrations
Cloth: 2104-X $18.95 *SPECIAL PRICE* $14.21
Paper: 2248-8 $12.95 *SPECIAL PRICE* $9.71

To order, mail this coupon or a copy to:

W. H. FREEMAN AND COMPANY

4419 West 1980 South, Salt Lake City, Utah 84104

☐ I enclose my check or money order made payable to W. H. Freeman and Company

☐ Visa ☐ MasterCard Expiration Date _____ Account # _____

Signature _____
(Credit card orders must be signed)

Name _____ Address _____
(Please print)

City/State/Zip _____

Quantity	Author	ISBN	Special Price	Total

Postage and handling *(Add $1.95 for the first book, $1.25 for each additional book.) $ _____

NY, CA, and UT residents add appropriate sales tax $ _____

TOTAL $ _____

*Canadian residents add $2.25 for postage and handling; $1.50 for each additional book. Please add 7% GST. Allow 4 weeks for delivery. All orders should be prepaid in U.S. dollars. 4897

The Mid-Ocean Ridge

It is the longest mountain chain, the most active volcanic area and until recently the least accessible region on the earth. New maps reveal striking details of how segments of the Ridge form and evolve

by Kenneth C. Macdonald and Paul J. Fox

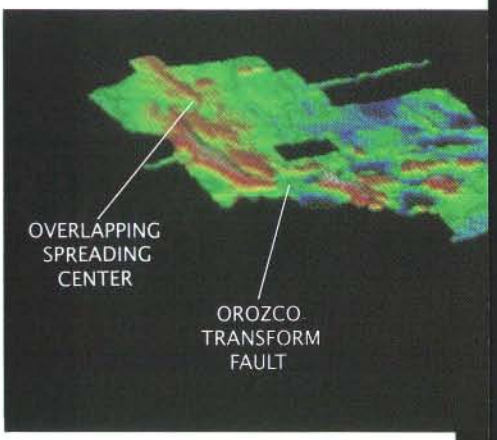
On July 8, 1982, we boarded the research vessel *Thomas Washington* to survey the East Pacific Rise, a volcanic mountain chain that lies under the Pacific Ocean. The Rise is part of the 75,000-kilometer-long formation known as the Mid-Ocean Ridge. Like the seam of a baseball, the Ridge winds around the globe from the Arctic Ocean to the Atlantic Ocean, around Africa, Asia and Australia, under the Pacific Ocean and to the west coast of North America. Even though the Ridge is by far the longest structure on the earth, less was known about its features than about the craters on the dark side of the moon.

Our colleagues from the Scripps Institution of Oceanography had recently equipped the *Thomas Washington* with a new type of sonar system, made by the General Instrument Corporation. Called SeaBeam, it could map a two-kilometer swath of the ocean floor in a single ping of the sonar. It would, we hoped, reveal the ocean floor in unprecedented detail, providing new insights into the forces that form and shape the Mid-Ocean Ridge.

After cruising southeast 2,500 kilometers from the Scripps marine facility in San Diego, we intersected the crest of the East Pacific Rise, located at a depth of about 2.5 kilometers. The Rise marks the boundary between the Pacific and Cocos tectonic plates, each a slab of the earth's crust and upper mantle. The plates separate at a rate of about 120 millimeters per year (twice the rate at which a fingernail grows). As the plates move apart, cracks form along the crest of the Rise, allowing molten rock to seep up from the mantle. Some of the molten rock overflows onto the ocean floor in tremendous eruptions. The magma then solidifies to form many square kilometers of new oceanic crust each year. Only a few kilometers above this activity, we felt like Lilliputians crawling along the spine of a slumbering giant that might awaken at any time.

As the SeaBeam probed the spine of this giant, we watched images of the seascape appear on monitors on board the *Thomas Washington*. We saw some familiar features: the elevated terrain that defines the axis of the Ridge and large breaks, called transform faults, that offset ridge segments by hundreds of kilometers. Yet we and Peter F. Lonsdale of Scripps also observed several unknown structures: segments that bend, ridges that overlap and oceanic crust that is warped and distorted near these features.

Since the early 1980s our colleagues in France, the U.K. and the U.S. have also surveyed many stretches of the East Pacific Rise as well as other parts of the Mid-Ocean Ridge. These efforts have revealed that the Ridge has many lateral discontinuities that partition its axis into segments. Although discontinuities differ in form and behavior, most of them are deeper and less active volcanically than the segments they define. As a result, the crest of the Ridge undulates up and down by hundreds of meters over distances of from 10 to 1,000 kilometers. During the past several years, we have come to understand how these discontinuities and segments evolve and how they are related to processes deep in the earth's crust and mantle.



American oceanographer Bruce C. Heezen aptly described the Mid-Ocean Ridge as "the wound that never heals." In 1956 he and W. Maurice Ewing noticed that the earthquakes in the ocean basin define a continuous belt encircling the world. Because the belt coincided with portions of the Mid-Ocean Ridge that were known at the time, they proposed that the earth was girdled by a continuous system of ocean ridges. Ever since their discovery oceanographers and geologists have tried to get a closer look at the Mid-Ocean Ridge to understand its origins.

The global geologic processes that form and shape the Ridge were not understood until 1960, when Harry H. Hess of Princeton University introduced the concept of seafloor spreading. Other workers further refined and developed his idea into the theory of plate tectonics. The theory posits that the crust and upper mantle are divided into a few dozen plates, such as the Pacific and the Cocos, which can move with respect to one another. If two plates separate, material from the mantle can well up, forming a ridge and new oceanic crust.

KENNETH C. MACDONALD and PAUL J. FOX have collaborated on many expeditions to the East Pacific Rise and the Mid-Atlantic Ridge. Macdonald is professor of marine geophysics at the University of California, Santa Barbara. In 1975 he received his Ph.D. in marine geophysics from the Massachusetts Institute of Technology and the Woods Hole Oceanographic Institution. Macdonald is drawn to the sea by his research, wind surfing and his wife, marine geologist Rachel M. Haymon. Fox is professor of oceanography at the University of Rhode Island. In 1972 he earned his Ph.D. in marine geophysics from Columbia University. When Fox and Macdonald are not cruising the seven seas, they wade in secluded mountain streams, where they try to catch unsuspecting trout with a well-presented fly.

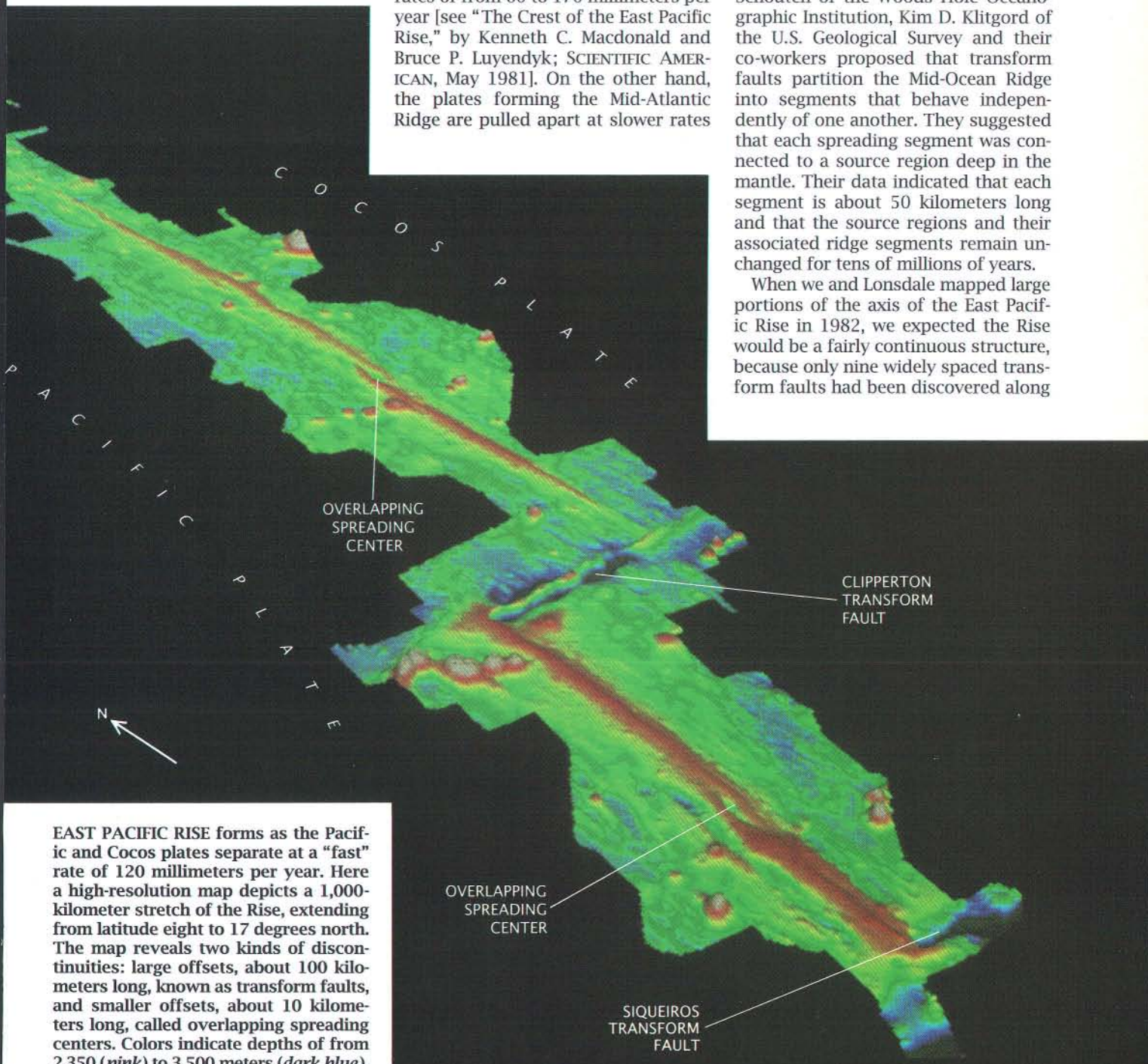
Reprinted from the June 1990 issue

The theory of plate tectonics accounts for the largest structures of the Mid-Ocean Ridge. Yet as early as 1960 H. William Menard of Scripps and Heezen discovered that the Mid-Ocean Ridge is a discontinuous structure. As they mapped the Ridge with sounding devices, they found several places where it was offset at right angles to its length. In 1965 J. Tuzo Wilson of the University of Toronto identified these discontinuities as transform faults: a boundary formed perpendicular to the length of the Ridge, where the edges of tectonic plates slide past each other in opposite directions. Later Richard N. Hey of the University of Hawaii realized the segments defined by two transform faults could shift in a direction parallel to the length of the Ridge. This type of discontinuity was called a propagating rift.

By the 1980s oceanographers had identified many transform faults and propagating rifts. They had also determined that different parts of the Mid-Ocean Ridge evolved at different rates. On the one hand, the plates forming the East Pacific Rise separate at "fast" rates of from 60 to 170 millimeters per year [see "The Crest of the East Pacific Rise," by Kenneth C. Macdonald and Bruce P. Luyendyk; SCIENTIFIC AMERICAN, May 1981]. On the other hand, the plates forming the Mid-Atlantic Ridge are pulled apart at slower rates of about 30 millimeters per year. Because of variations in spreading rates and the rate at which magma is supplied to ridges, the topography of fast-spreading ridges differs from that of slow-spreading ones. The crest of a fast-spreading ridge is defined by an elevation of the oceanic crust several hundred meters high and five to 20 kilometers wide. In contrast, the axis of a slow-spreading ridge is characterized by a rift valley a few kilometers deep and about 20 to 30 kilometers wide.

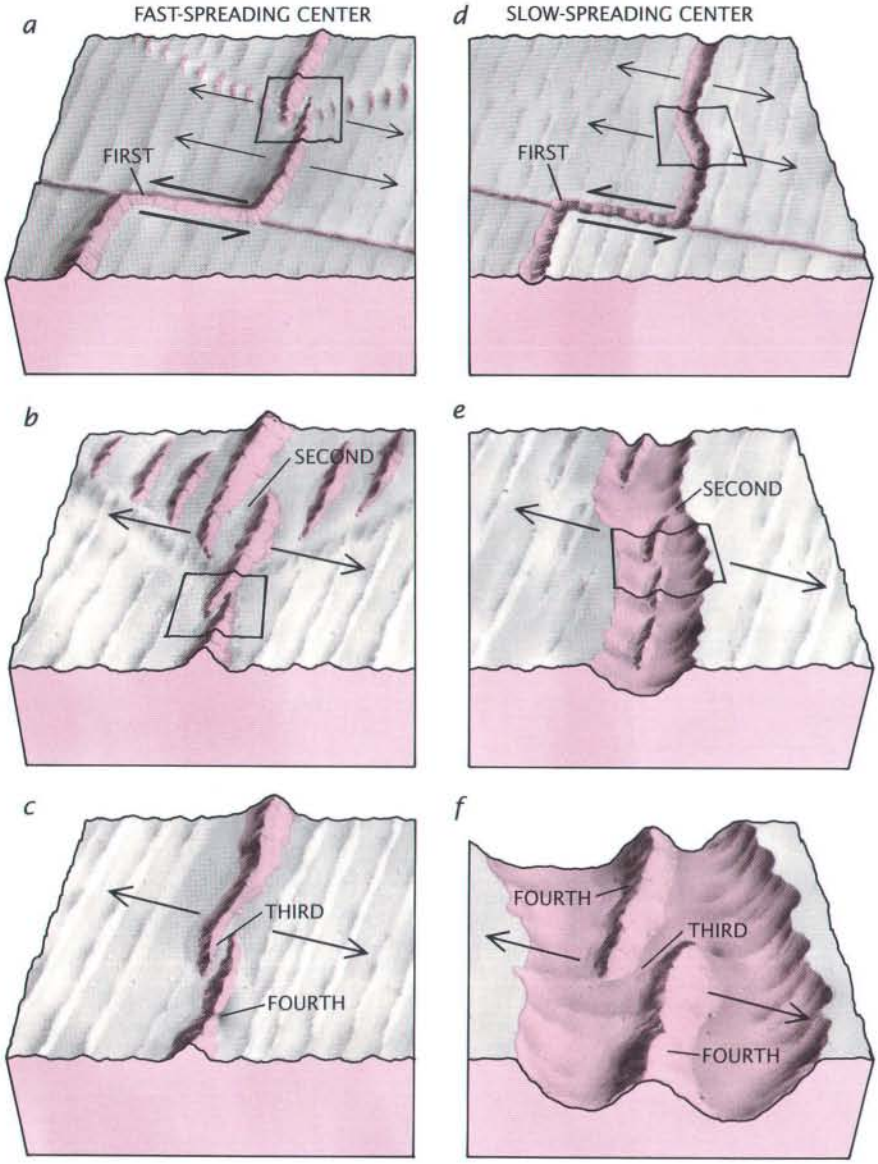
In the early 1980s, based on observations of the Mid-Atlantic Ridge, Hans Schouten of the Woods Hole Oceanographic Institution, Kim D. Klitgord of the U.S. Geological Survey and their co-workers proposed that transform faults partition the Mid-Ocean Ridge into segments that behave independently of one another. They suggested that each spreading segment was connected to a source region deep in the mantle. Their data indicated that each segment is about 50 kilometers long and that the source regions and their associated ridge segments remain unchanged for tens of millions of years.

When we and Lonsdale mapped large portions of the axis of the East Pacific Rise in 1982, we expected the Rise would be a fairly continuous structure, because only nine widely spaced transform faults had been discovered along



EAST PACIFIC RISE forms as the Pacific and Cocos plates separate at a "fast" rate of 120 millimeters per year. Here a high-resolution map depicts a 1,000-kilometer stretch of the Rise, extending from latitude eight to 17 degrees north. The map reveals two kinds of discontinuities: large offsets, about 100 kilometers long, known as transform faults, and smaller offsets, about 10 kilometers long, called overlapping spreading centers. Colors indicate depths of from 2,350 (*pink*) to 3,500 meters (*dark blue*).

its 5,000-kilometer length. To our surprise, the axis of the Rise was frequently disrupted by many small offsets (more than 40 have been mapped to date). These discontinuities partitioned the Ridge into segments ranging in length from 10 to 200 kilometers. Unlike transform faults, these offsets were characterized by overlapping ridge tips, and they did not have a clearly defined fault that connected the tips [*see illustration below*]. Since their discovery we have mapped the off-axis regions around these overlapping offsets and have learned that the features evolve rapidly. In addition, we have found that the discontinuities can migrate along the Ridge, at varying speeds and in various directions. Individual segments bounded by those discontinuities can apparently lengthen or shorten. High-resolution maps have also documented similar nonrigid discontinuities on the slow-spreading Mid-Atlantic Ridge.



DISCONTINUITIES in the Mid-Ocean Ridge can be classified according to shape, size and longevity. For a fast-spreading center, such as the East Pacific Rise, a first-order discontinuity (*a*) is a transform fault, where rigid plates slide past each other. It offsets the Ridge by at least 50 kilometers. A second-order discontinuity (*b*) is usually a large overlapping spreading center that offsets the Ridge by at least two kilometers. A third-order discontinuity (*c*) is a small overlapping spreading center that offsets the Ridge by 0.5 to two kilometers. A fourth-order discontinuity (*c*) is characterized by slight deviations in axial linearity. For a slow-spreading center, such as the Mid-Atlantic Ridge, a first-order discontinuity (*d*) is also typically a transform fault, but it represents a break in a rift valley rather than a ridge crest. A second-order discontinuity (*e*) is a bend, or jog, in the rift valley. A third-order discontinuity (*f*) is a gap between chains of volcanoes, whereas a fourth-order discontinuity (*f*) is a small gap within a chain of volcanoes. First- and second-order structures are usually flanked by distorted crust that formed as the discontinuity evolved. They are known to persist longer than third- and fourth-order discontinuities because the oceanic crust near the higher-order structures does not show evidence of distortion.

To determine the origin of these discontinuities, we and our colleagues attempted to find connections between segmentation and volcanic activity. Although volcanism can change greatly from one segment to the next, it does vary systematically along the length of each segment. The least active regions are deep discontinuities, whereas the most active regions are shallow centers of segments [see "The Oceanic Crust," by Jean Francheteau; SCIENTIFIC AMERICAN, September 1983].

From these observations and others, we, Schouten and our colleagues developed a magma-supply model of ridge segmentation. In the mantle at a depth of from 30 to 60 kilometers, rocks are heated to high temperatures, but because they are usually subjected to high pressure, they remain in a solid state. The environment is somewhat different at the boundary between tectonic plates. As plates separate, some of the rock decompresses and melts. The molten rock then percolates up through the mantle and fills a shallow chamber in the crust beneath the crest of the ridge. As the chamber swells with magma and begins to expand, the crest of the ridge can be pushed upward by the buoyant forces from both the molten rock in the magma chamber and the broader region of hot rock in the upper mantle [*see illustration on page 146*].

According to the magma-supply model of segmentation, the greater the supply of molten and hot rock to a region, the higher the overlying ridge segment will be elevated. Furthermore, the rate and volume of the molten-rock supply can change from region to region, creating variations in the morphology of the different overlying segments.

The magma-supply model also accounts for smaller structural variations. As magma in the chambers migrates laterally along the ridge axis, the thin, brittle crust above the magma chamber stretches and fractures. The magma can erupt through these fractures to the ocean floor. As the cracks continue to grow, volcanic eruptions follow in their wake. The eruptions will continue until the production of mag-

ma subsides and the supply of magma is exhausted. Temporal variations in melt delivery affect a segment's evolution: when a segment is well supplied with molten rock as compared with its neighbors, the segment tends to lengthen, and when it is poorly supplied, the segment shortens. It is this swelling and shrinking of the magma-supply system, in response to plate separation, that initiates the lengthening or shortening of segments and the migration of small discontinuities.

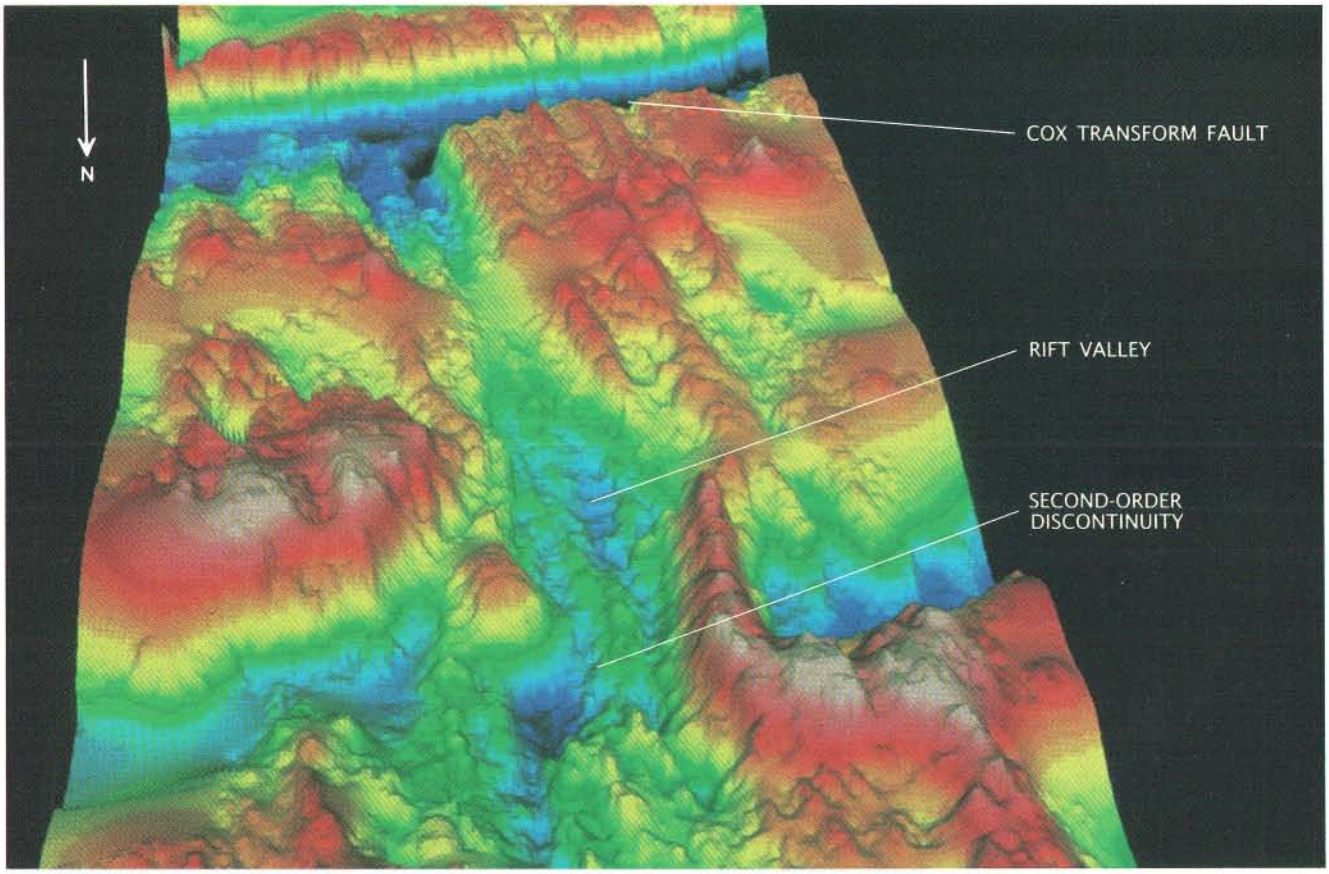
The magma-supply model appears to agree with seismic and gravitational measurements of the East Pacific Rise. Seismic measurements reveal that a good reflector of sound energy exists about 1.2 to 2.5 kilometers beneath the shallow portions of each ridge segment. This reflector often deepens and then disappears near discontinuities. In 1987 Robert S. Detrick of the University of Rhode Island and his co-workers proposed that the reflector is the roof of a magma chamber. The strength of sound reflection can be explained by a thin cap of nearly 100 percent melt along the top of the chamber.

Most geologists and oceanographers now agree the reflector is a long, shallow body of magma beneath the ridge surrounded by hot rock. John A. Orcutt of Scripps and his colleagues have made seismic measurements along the northern East Pacific Rise, which suggest such a chamber of molten rock is only two to four kilometers wide and less than one kilometer thick. The magma chamber is surrounded by a wider region of very hot (perhaps slightly molten) rock. The reservoir may be six to 10 kilometers wide and three to six kilometers thick. This region of hot rock extends at least to the base of the oceanic crust and probably a few kilometers into the upper mantle [*see illustration on page 149*].

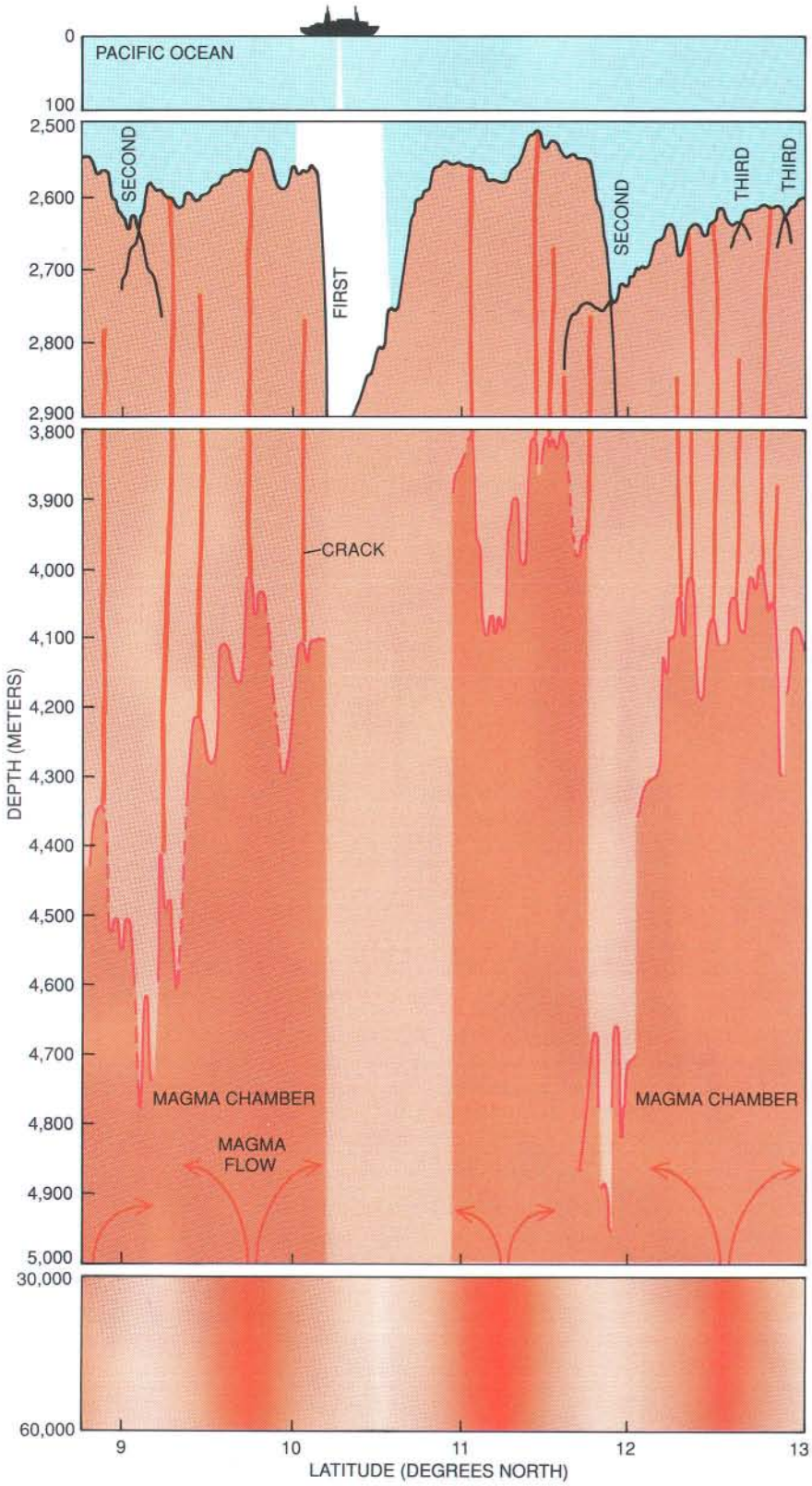
The presence of magma chambers and hot-rock reservoirs has been supported by precise measurements of the gravitational field there, which indicate the presence of a buoyant mass beneath the ridge axis. From both seismic and gravitational measurements, workers have deduced that the magma chamber resembles a mushroom in cross section: it has a narrow stalk of partial melt feeding a wide but very thin lens of pure melt.

Seismic evidence has not definitively proved that magma chambers exist beneath slow-spreading formations such as the Mid-Atlantic Ridge. Other measurements, however, seem to support a magma-supply model for slow-spreading ridges. Donald W. Forsyth and Ban-Yuen Kuo of Brown University and Jian Lin and G. Michael Purdy of Woods Hole found anomalies in the gravitational field, which were centered over the shallowest parts of several segments of the Mid-Atlantic Ridge. The best explanation for these anomalies is an upwelling of hot mantle material or a thickening of the oceanic crust beneath the shallow parts of each segment. Both interpretations are consistent with the magma-supply model.

It was a great relief that the seismic and gravitational measurements supported, at least in a general sense, the magma-supply model of segmentation. We and many other tectonicists and geochemists had stuck our necks out fairly far with that hypothesis. True, some of us thought the chamber would



MID-ATLANTIC RIDGE emerges as the South American and African plates pull apart at the "slow" rate of approximately 30 millimeters per year. The axis of the Ridge is marked by a two-kilometer-deep rift valley, which is typical of most slow-spreading ridges. The map reveals a 12-kilometer jog of the rift valley, a second-order discontinuity, and also shows a first-order discontinuity called the Cox transform fault. Colors indicate depths of from 1,900 (*pink*) to 4,200 meters (*dark blue*).



MAGMA seeps up from deep within the mantle to form the East Pacific Rise (shown in cross section along the crest of the ridge). Investigators speculate that partially melted rock from depths of 30,000 to 60,000 meters percolates upward and is produced in greater quantities in some areas (*dark red*) than in others (*light red*). They propose that the molten rock fills and expands magma chambers. Seismic measurements suggest that the tops of the chambers are at the depth indicated by the broken red line. Molten rock ascends from the magma chamber through cracks in the crust and then solidifies or erupts onto the ocean floor. The depth of the ridge (*black line at top*) was determined from sonar measurements. The chamber breaks below discontinuities of order one, two and sometimes three.

be larger, and it remains to be seen if magma actually flows laterally below the ridge axis, but significant evidence has been found to support the model.

The magma-supply model has been quite successful in accounting for the many different types of discontinuities and segments. Such structures are classified as first, second, third or fourth order according to their size, longevity, geometry and behavior. It has been demonstrated that first-, second- and third-order structures are fundamental components of both fast- and slow-spreading ridges. (The role of fourth-order features remains unsolved.) Because these structures have been investigated in more detail on fast-spreading ridges, we will describe them in that setting first.

The most common type of first-order discontinuity is the transform fault. It appears where rigid plates slide past each other. First-order discontinuities offset the ridge segments by at least 20 kilometers and usually more than 50 kilometers. Hence, most transform faults were large enough to be revealed by early reconnaissance-mapping efforts. These discontinuities typically define segments from 200 to 800 kilometers long.

On the ocean floor, transform faults appear to be narrow, straight bands linking the ends of segments. These bands can be traced in the flanks of a ridge for hundreds to thousands of kilometers [*see illustration on page 143*]. Such traces indicate that first-order structures persist for millions to tens of millions of years.

A first-order segment can be broken up by several second-order discontinuities that are usually spaced from 50 to 300 kilometers apart. Unlike first-order structures, however, second-order segments are not rigid, and their motion is not concentrated along a narrow fault zone. Second-order discontinuities are complex features characterized by oblique and overlapping structures.

Second-order discontinuities are typically features that resemble the arms of two people who are preparing to shake hands. The arms (ridges) are extended in such a way that the hands (the curved ends of ridges) overlap. The distance between the "hands" varies from one to 20 kilometers. The offset is typically three times shorter than the distance that the ridges overlap. Such a feature is known as an overlapping spreading center [*see illustration on opposite page*].

When overlapping spreading centers were discovered in 1982, we could not

account for many of their characteristics. Why did so many centers have an overlap-to-offset ratio of 3 to 1? What happened to the crust that lies between the overlapping ridges? Why did the ridges create a distinctive curving pattern?

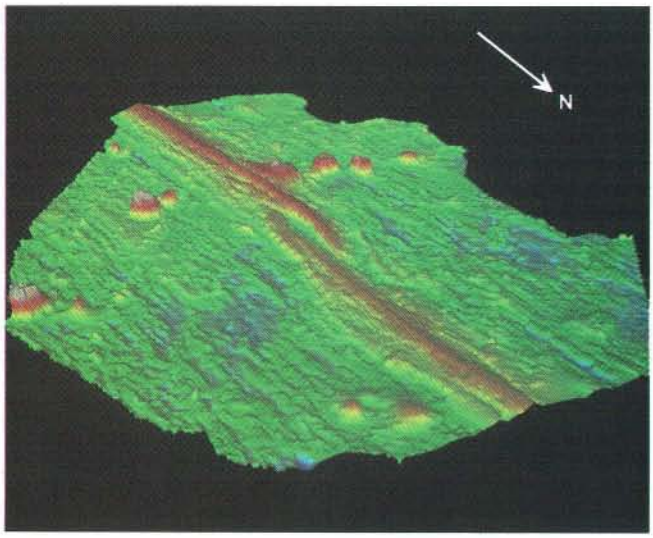
In 1984 David D. Pollard of Stanford University, Jean-Christophe Sempere, then at the University of California at Santa Barbara, and one of us (Macdonald) found that the highly repetitive shape of overlapping spreading centers could be explained by the way cracks develop and propagate along ridges. As tectonic plates are pulled apart, cracks form perpendicular to the direction of tension. In the middle of a segment the direction of stress is usually perpendicular to the ridge axis, so the cracks will lengthen parallel to the ridge. In the region of overlapping segments, however, the direction of stress can vary. As a crack from the middle of a segment begins to grow toward the region of overlap, the crack first deflects away from the region and then hooks toward it [*see illustration on next page*]. The crack allows magma to erupt onto the ocean floor, and a new ridge tip is formed. But once the cracks overlap by a distance that approaches three times their offset, the crack propagation stalls abruptly. Soon after, a new crack begins to develop behind the first. As the second develops, the first ridge tip is shed off onto the flanks because of plate separation.

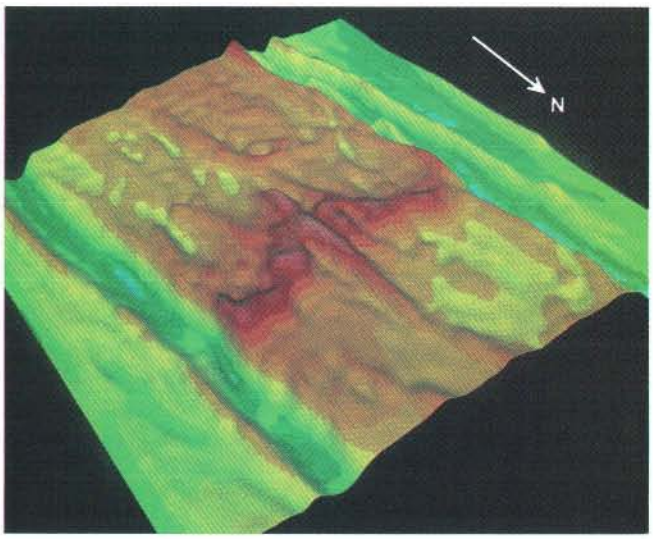
Spreading centers that overlap by more than several kilometers usually leave "wakes" of deformed oceanic crust up to 80 kilometers wide. The ocean floor within such a disturbed region, called a discordant zone, is 100 to 300 meters deeper than the surrounding ocean floor; in like manner, the overlapping spreading centers lie 100 to 300 meters deeper than the shallow, magmatically robust parts of the ridge segments. These features have emerged from maps that several expeditions have made of the flanks of the East Pacific Rise. The maps of the discordant zones also show curved fossil ridge tips 10 to 40 kilometers long, which have been cut off at overlapping spreading centers.

The magma-supply model appears to account for the structure of the overlapping spreading centers. It seems that overlapping spreading centers are at the ends of magma sources and tend to be deprived of magma. If this is true, the crust created at overlapping spreading centers may be up to 50 percent thinner than the six-kilometer-thick crust near the centers of each segment. Detailed seismic and gravitational measurements need to be carried out in these areas to test this idea.

Measurements of the earth's magnetic field at overlapping spreading centers support the idea that such centers occur where the magma supply is low. It turns out that lava that erupts from small magma chambers, which alternately solidify and become replenished, tends to contain more iron-rich minerals in a highly magnetized state. On the other hand, magma chambers large enough to remain molten between episodes of magma replenishment produce lava that is magnetically weak. Because rock near overlapping spreading centers is often much more strongly magnetic than rock elsewhere along the ridge, it seems likely that the centers are fed discontinuously from isolated pockets of magma.

Based on the age of the crust into which the discordant zones extend and on the patterns of off-axis wakes, Laura J. Perram, Suzanne M. Carbotte and Marie-Helene Cormier of the University of California at Santa Barbara have demonstrated that second-order segments persist as discrete entities for up to several million years. The discontinuities may slowly oscillate in position by 10 to 20 kilometers on the ridge or may migrate along the ridge many tens of kilometers at rates of 20 to 100 millimeters per year. A discontinuity tends to move in spurts; a ridge





OVERLAPPING SPREADING CENTER, which cuts across the East Pacific Rise near latitude 12 degrees north, was surveyed to determine its topography (*left*) and magnetization (*right*). The topographical map shows that the overlapping spreading center offsets the Rise by eight kilometers. Colors indicate depths of from 2,350 (*pink*) to 3,500 meters (*dark blue*). The two arms of the discontinuity overlap by 27 kilometers. The arms narrow and deepen near the discontinuity, presumably because the supply of magma to the region is low. The ocean floor near the discontinuity—also known as the wake—is unusually deep and is littered with ridge tips, especially on the west side. It turns out that regions that are not well supplied with magma are highly magnetized. In the map at the right, magnetization decreases in strength from red to yellow regions. The map reveals the wake (*red*) of the overlapping spreading center. The green-blue troughs were created 700,000 years ago when the earth's magnetic field reversed polarity. The wake shows that the overlapping spreading center emerged about 700,000 years ago, migrated north a short distance and then moved slowly south at 70 millimeters per year. In the past 200,000 years migration to the south has accelerated to 200 millimeters per year.

segment can lengthen at rates of several hundred millimeters per year but then may retreat and shorten for a time before making the next surge forward. In this way, the ridge tips at a second-order discontinuity appear to be "dueling" as they surge back and forth along the ridge, generally making slow progress in either direction [*see illustration below*].

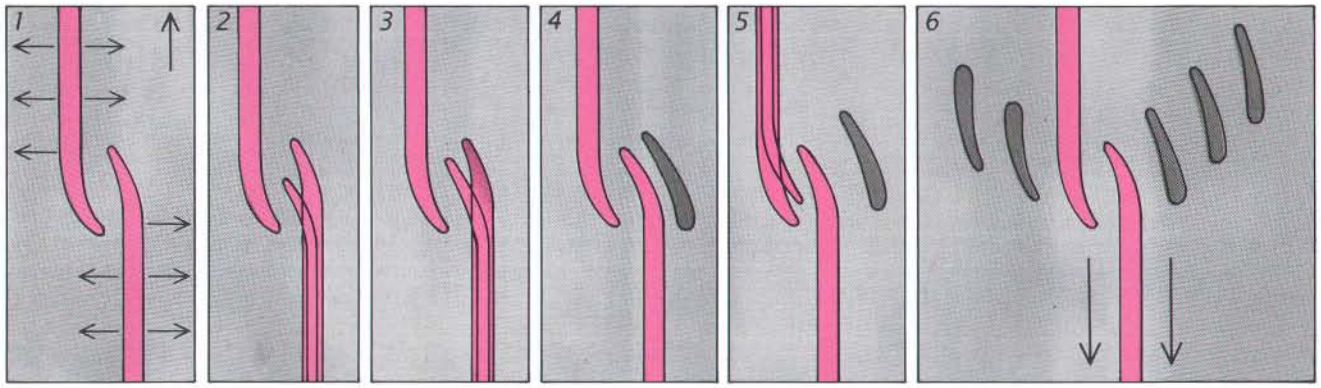
Along the East Pacific Rise, third-order discontinuities usually consist of overlapping spreading centers that offset the ridge by less than three kilometers. Segments defined by third-order discontinuities are from 30 to 100 kilometers long. Third-order discontinuities have been shown to correspond with breaks in magma chambers.

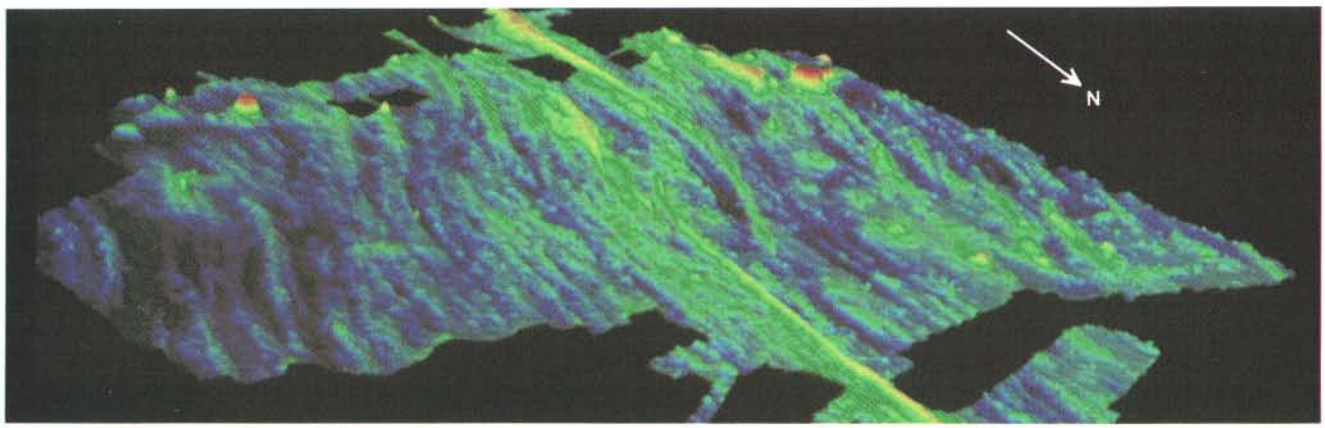
The ridge segments defined by third-order discontinuities leave little or no evidence of off-axis wakes. Because they produce little trace in old oceanic crust on the ridge flanks, we can conclude that third-order discontinuities are geologically short-lived. In fact, we estimate that they are younger than 10,000 years—the time it takes a fast-spreading ridge to generate two kilometers of crust.

Fourth-order discontinuities are either subtle bends or tiny offsets less than 500 meters in size. The structures are often called DEVALS (for slight DEViations in Axial Linearity). DEVALS are usually spaced from 10 to 40 kilometers apart. A DEVAL may be the manifestation of a single major eruption and therefore may be as young as hundreds to thousands of years old.

DEVALS are very difficult to detect. They can barely be resolved with sonar systems such as SeaBeam, and seismic measurements are not much help either. In some cases, the magma chamber beneath a DEVAL deepens slightly and, in rare instances, exhibits an apparent break. In most cases, the chambers below fourth-order discontinuities are fairly continuous. During the 1982 cruise, one of us (Fox) pointed out to the other (Macdonald) that he had found several DEVALS in the SeaBeam maps. Macdonald then told Fox that he had been staring at the maps too closely on a rolling ship. We soon agreed that we should focus on the larger offsets if we wanted people to believe our ideas.

Indeed, fourth-order segments (the sections of ridge between DEVALS) were not recognized as distinct and significant features until 1986, when Charles H. Langmuir of the Lamont-Doherty





OFF-AXIS FEATURES are generated by an overlapping spreading center, as illustrated in the diagram (*top*) and the map (*bottom*). An overlapping spreading center is depicted (*1*). A crack develops to the south of the eastern ridge tip (*2*), allowing molten rock to surface and form a new tip. The new tip lengthens until it overlaps the western ridge by three times the distance that separates them (*3*). As the regions of rock continue to pull apart, the original eastern ridge tip breaks off and migrates away (*4*). A new western tip begins to form (*5*). After many episodes of ridge-tip formation and migration (*6*), the off-axis structures show a net migration to the south. The high-resolution map of a region near 21 degrees south reveals an overlapping spreading center that offsets the East Pacific Rise by 12 kilometers. The discontinuity has had a complex evolution during the past two million years. Migration rates have exceeded 200 millimeters per year as northern and southern ridge tips have surged back and forth, but net migration toward the south has averaged 20 millimeters per year. Numerous abandoned ridge tips within a wake of unusually deep seafloor can be seen on both sides of the overlapping spreading center. Seafloor structure is disrupted across an 80-kilometer-wide swath adjacent to this discontinuity. Colors indicate depths of from 2,350 (*pink*) to 2,900 (*yellow*) to 3,500 meters (*dark blue*).

Geological Observatory, John F. Bender of the University of North Carolina at Charlotte and their colleagues analyzed the geochemistry of a 500-kilometer stretch of the East Pacific Rise. The workers collected rock samples from precise sites on the ocean floor to see if structural segmentation could be associated with variations in rock chemistry. They found that the rocks within each fourth-order segment had a similar composition, whereas rocks from other segments had different chemical signatures. Such measurements have documented the fundamental nature of segmentation over a range of scales and have helped to identify many other DEVALS.

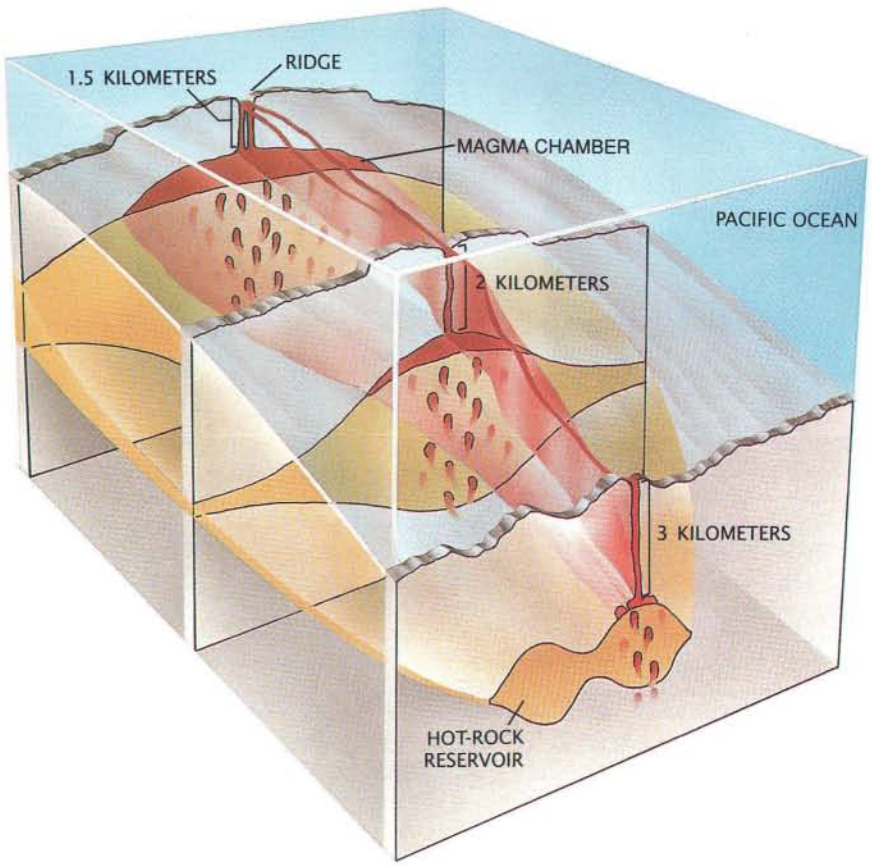
Why do DEVALS differ in chemical composition? One theory posits that small blockages divide the magma chamber beneath adjacent fourth-order segments. These divisions would prevent the mixing of the magma in the chambers. Yet only a few such breaks have been detected in magma chambers beneath DEVALS.

Another theory suggests that small batches of molten rock from isolated sources in the upper mantle may be injected locally into a magma chamber and may erupt onto the ocean floor before much mixing occurs. This process would create a fourth-order segment with a distinctive rock chemistry. More evidence is needed to confirm or refute these ideas.

We, Carbotte and Nancy R. Grindlay of the University of Rhode Island have documented several kinds of first-, second- and third-order discontinuities at slow-spreading ridges in the South Atlantic. Like the first-order discontinuities on fast-spreading ridges, the slow-spreading counterparts are transform faults. Second-order discontinuities on slow-spreading ridges are defined by a lateral step of the rift valley or by a deep oblique basin along which the offset rift valleys are linked. The second-order discontinuities persist for millions of years—longer on average than second-order discontinuities on fast-spreading centers. Second-order discontinuities on slow-spreading centers also migrate more slowly along the ridge axis than their fast-spreading counterparts. Third-order discontinuities at slow-spreading centers are small offsets in long volcanic chains within the rift valley floor, whereas fourth-order discontinuities may be small gaps between volcanoes.

Oceanographers, tectonicists and geochemists have just begun to understand some of the implications of segmentation for both slow- and fast-spreading ridges. We have found clear examples of first-, second-, third- and fourth-order structures and everything in between. Do segments evolve from fourth through first order and back again? We know that segmentation has been a fundamental process for at least 100 million years. Has segmentation played a role over a much longer period? Investigators have studied exotic faunal communities that flourish near hot springs on the Mid-Ocean Ridge. Can the survival and migration of these communities be linked to the longevity of a given segment?

These questions will be the focus of research for a program called the Ridge Interdisciplinary Global Experiments (RIDGE). Among the many goals of the program are to map the axis and flanks of the entire Mid-Ocean Ridge and to generate more detailed images of off-axis features. Even today geologists and oceanographers have mapped less than 5 percent of the seafloor. More than half of the earth's crust remains to be explored.



MAGMA CHAMBER is thought to extend below fast-spreading ridges. The magma chamber is a lens of mostly molten rock. The chamber sits atop a reservoir of partially melted rock. The chamber and reservoir are small and poorly supplied with molten rock near a discontinuity (*deep region in foreground*). Yet they can be larger and well supplied at a distance away from the discontinuity (*background*).

FURTHER READING

THE GEOLOGY OF DEEP-SEA HOT SPRINGS. Rachel M. Haymon and Ken C. Macdonald in *American Scientist*, Vol. 73, No. 5, pages 441–449; September/October 1985.

SEGMENTATION OF MID-OCEAN RIDGES. Hans Schouten, Kim D. Klitgord and John A. Whitehead in *Nature*, Vol. 317, No. 6034, pages 225–229; September 19, 1985.

PETROLOGICAL AND TECTONIC SEGMENTATION OF THE EAST PACIFIC RISE, 5°30'–14°30'N. Charles H. Langmuir, John F. Bender and Rodey Batiza in *Nature*, Vol. 322, No. 6078, pages 422–429; July 31, 1986.

MULTI-CHANNEL SEISMIC IMAGING OF A CRUSTAL MAGMA CHAMBER ALONG THE EAST PACIFIC RISE. R. S. Detrick, P. Buhl, E. Vera, J. Mutter, J. Orcutt, J. Madsen and T. Brocher in *Nature*, Vol. 326, No. 6108, pages 35–41; March 5, 1987.

A NEW VIEW OF THE MID-OCEAN RIDGE FROM THE BEHAVIOUR OF RIDGE-AXIS DISCONTINUITIES. Ken C. Macdonald, P. J. Fox, L. J. Perram et al. in *Nature*, Vol. 335, No. 6187, pages 217–225; September 15, 1988.

The Transistor

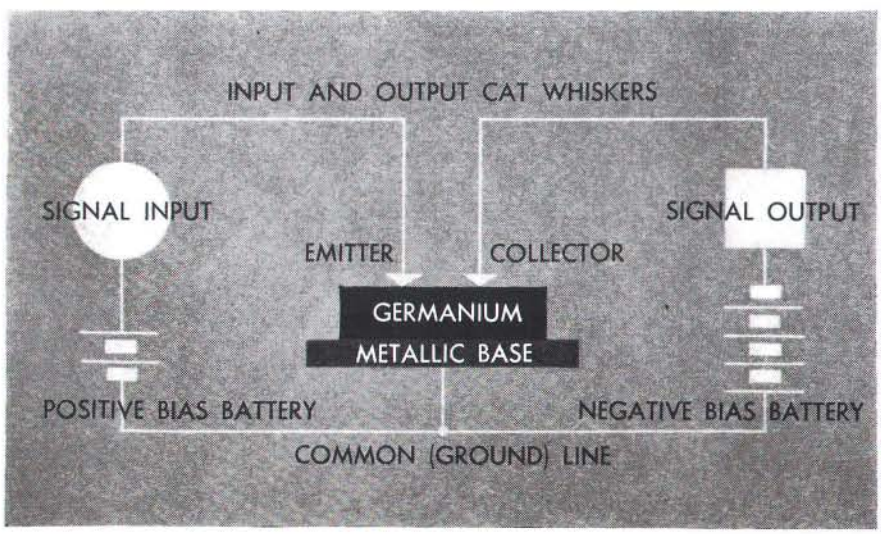
Basic research in the electrical properties of solids has opened up an entirely new way of manipulating electrons to do useful work

by Frank H. Rockett

In 1906 a young American electrical engineer named Lee De Forest discovered that if an electrified wire grid was placed across the path of a stream of electrons in a vacuum tube, the flow of electrons could be controlled in some rather interesting ways. The flow could be interrupted, reduced or stopped entirely; a feeble current of electrons entering at one end of the tube could be "amplified" to a powerful current at the outgoing end. It was this classically simple invention by De Forest that gave birth to the tremendous technology of electronics. From it came radio, television, radar, X-ray cameras, electron microscopes, guided missiles, electronic calculators, robot machine tenders, electronic burglar alarms, instruments that examine materials for invisible flaws, doors that open themselves—and doubtless greater wonders are yet to come. The electronic tube is easily one of the most ingenious inventions and most versatile tools of our fabulous century.

FRANK H. ROCKETT, an electrical engineer, was associate editor of the journal *Electronics* when he wrote this article in 1948.

Since De Forest's elementary discovery, the electronic tube has been developed enormously. Electronic theory also has advanced rapidly, and it now appears that the vacuum tube is far from the last word. Within the past few months, a group of physicists at the Bell Telephone Laboratories has made another profound and simple finding, which may rank in importance with that of De Forest. In essence, it is a method of controlling electrons in a solid crystal instead of in a vacuum. This discovery has yielded a device called the transistor (so named because it transfers an electrical signal across a resistor), which can do many of the things that a vacuum tube does. Indeed, it has certain advantages over the vacuum tube. It reduces the complicated, delicate tube to a simple rig consisting basically of a couple of fine wires—cat's whiskers in the radioman's language—and a small crystal; no vacuum is needed. The transistor does not need to heat up, as a vacuum tube does, and so it goes to work instantly. It operates on a tiny amount of power—about one tenth of that used by an ordinary flashlight bulb. And it can be made almost vanishingly small. The present experimental model is about the size of the eraser on the end of a pencil.



SIMPLE CIRCUIT uses a transistor in place of a vacuum tube. When a small signal passes over the surface of the transistor between the cat's whiskers, a larger current through the transistor is modulated in replica of the small signal.

The technological fruits of this invention already appear extensive. The size of vacuum tubes is an important consideration in electronics, for it largely determines the size of the apparatus in which they are used. A television receiver requires about two dozen tubes; the celebrated computing machine at the University of Pennsylvania known as ENIAC has 18,000. With ingenuity and painstaking labor, "subminiature" vacuum tubes only an inch long have been produced for some special purposes, but the transistor promises to reduce electronic equipment in general to an even smaller scale. Not only is the transistor itself tiny, but it needs so little power and uses that little so efficiently (as a radio amplifier its efficiency is 25 percent, against a vacuum tube's 10 percent) that the size of batteries needed to operate portable devices can also be reduced. Thus, the transistor makes possible tinier hearing aids, really small portable radios, more compact electronic devices for aircraft and a great reduction in the bulk of stationary equipment. In combination with printed circuits—the compact new wiring system—it may open up entirely new applications for electronics. The transistor also suggests the possibility of a considerable improvement in telephone transmission, because amplifiers for long-distance cables can be built small and mounted inconspicuously on telephone poles, and a miniature amplifier may even be built into the telephone receiver to strengthen weak signals.

Reprinted from the September 1948 issue

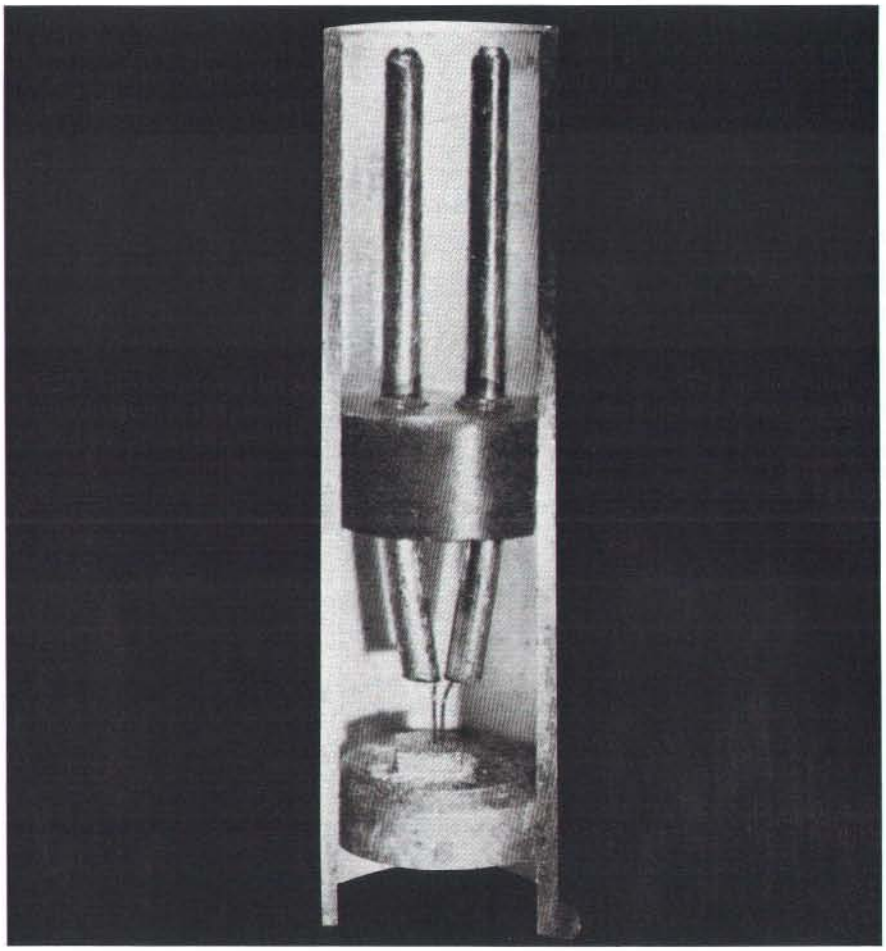
Beyond this, the transistor vastly simplifies the manufacture and maintenance of electronic equipment. Because of its simple, sturdy construction, it will be longer-lived and possibly less costly than vacuum tubes. The transistor has important limitations: its power output in the present research stage is small (a maximum of about one fortieth of a watt), and the highest frequency at which it can operate is about 10 megacycles (10 million cycles per second). But its power and frequency range are sufficient for most purposes in the regular broadcast, television and short-wave regions of the radio spectrum.

The transistor is the unexpected product of purely scientific curiosity. To understand how it was conceived and how it works, one must examine the functions of the electronic tube and the way in which electric current is conducted by a solid. The basic purposes of an electronic tube are to convert an alternating current into a direct current (called rectification), to amplify the signal, to break it up into pulses instead of a continuous wave, or to make it oscillate, that is, beat in a regular rhythm at a calculated frequency. The tube itself was invented in 1905 by the English physicist J. Ambrose Fleming, who observed that if an alternating current was passed to a filament inside a vacuum tube, the electrons would boil off the end of the filament as free particles and would travel across the vacuum to a positively charged plate at the other end. (This phenomenon, known as the Edison effect, had been noticed much earlier by Thomas Edison, but he had been unable to explain it and had made no practical use of it.) As long as the current was kept on, the electrons would move only toward the attracting positive plate; hence, the tube was an easy means of changing alternating current into a direct, one-way signal. Fleming's tube, called a diode because it had two electrodes—the filament and the plate—could be used as a detector for radio signals. But De Forest's addition of the grid to control the electrons, making the tube a triode, was the step that gave the tube its great versatility and usefulness. Now the signal could be controlled and amplified (since a small number of electrons on the grid governed the flow of a much larger number from the filament to the plate). It could also be modified in other ways.

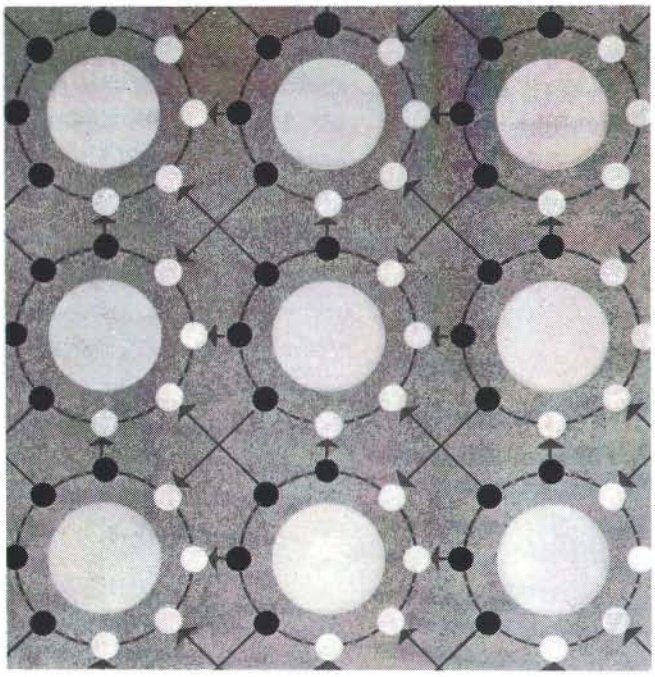
When not in a vacuum, electrons obviously are much less easy to control, since they cling more or less firmly to orbits about the nuclei of atoms. Whether a solid will conduct electricity depends on the degree of freedom of its electrons. Copper, a good conductor, has a single electron in its outer orbit or shell, and this relatively free electron readily serves to carry current. Most metals have such loosely held electrons and hence are good conductors. On the other hand, an element such as sulfur, whose electrons are all locked in place by tight bonds with the nucleus and with other atoms, does not conduct electricity; it is an insulator.

Between these extremes there is a class of materials known as semiconductors that furnish an occasional free electron for carrying current. Silicon and germanium are examples; they have about one free electron for every 1,000 atoms (as contrasted with copper, which has one for every atom). These semiconductors have long possessed a special interest for electronic researchers. The important fact about them is that the number of current-carrying electrons in them can be controlled. They can be made to act as conductors under some conditions and as insulators under others. Indeed, they are so sensitive that the current flowing in a semiconductor can be controlled by the brightness of a light shining on it in a region where a fine wire touches it. So this class of materials has been adapted to many uses. The crystal detector, which was used in early radios and is now employed in an improved form in radar sets, is a semiconductor.

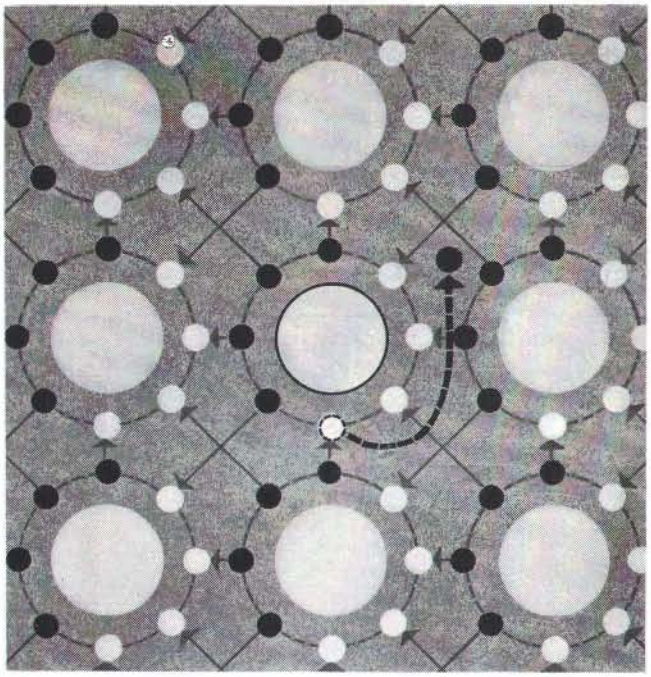
It was research into some of the mysterious electrical properties of semiconductors that led to the development of the transistor. It is helpful to try to visualize the electrical behavior of one of these substances. Picture a crystal of silicon (or germanium), which has four electrons in its outer shell—so-called valence electrons that hook the atoms together. Because they are fully occupied in forming bonds between the atoms, the electrons are not available for carrying electricity. Now suppose some impurity which has five valence electrons, say, an atom of phosphorus, gets into the crystal. Four of these electrons become busy form-



EARLY TRANSISTOR was about half the size of a paper clip. The metal tube of the transistor is cut away. The germanium crystal is the tiny block on the disk at bottom. Two cat's whiskers are mounted on heavy leads.



ATOMS OF GERMANIUM, shown here in a schematic crystal lattice, have four electrons (*black dots*) in outer orbits. These move freely into gaps in adjacent orbits.



ATOM OF PHOSPHORUS, introduced into a germanium crystal, has one more outer electron. This is free to travel through the crystal, improving its conductivity.

ing bonds with silicon atoms, but the fifth is free to carry current.

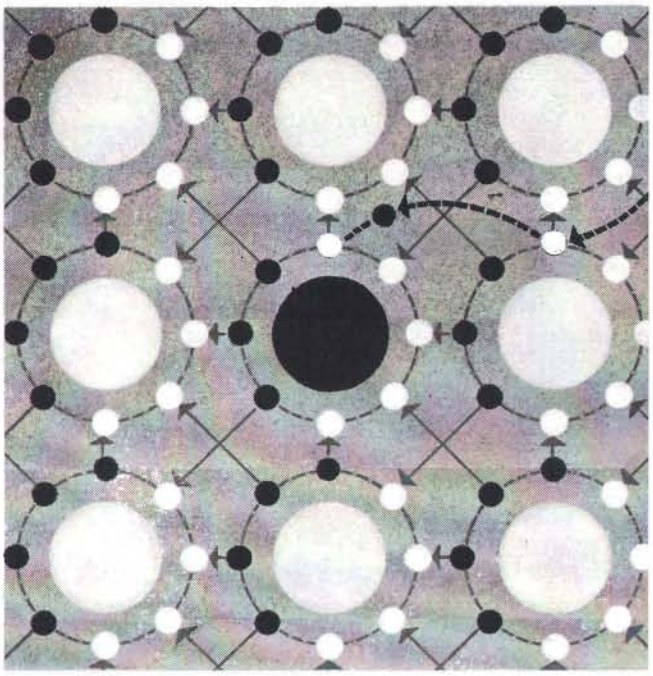
A more interesting case, and the one with which we are chiefly concerned here, is an impurity with three valence electrons, such as boron. One of the bonds needed for union with the silicon atoms is missing. The result is a state of disequilibrium, as the physicists say; there is some shifting around of bonds, but however they arrange themselves there is bound to be a missing electron. Because it is much easier to consider the movements of the gap created by the single missing electron than to follow the movements of the numerous other electrons as they create and fill in gaps, the missing electron is treated as an actual physical entity, although it is called a "hole." It has all the properties of an electron, such as mass and charge, except that, being the absence of an electron, its charge is positive instead of negative.

This, then, is a rough picture of the theory: the ability of a crystal semiconductor to conduct electricity is due to the presence of impurities that free some of the electrons which would otherwise be occupied in linking atoms. But a physicist at the Bell Telephone Laboratories, John Bardeen, became curious about a phenomenon that seemed to leave a hole in the theory. When a semiconductor is placed between two metallic contacts in an electrical circuit, one of the contacts being the point of a fine wire and the other a metal block, the arrangement acts as a rectifier, in a manner somewhat similar to the electronic tube. The reason is that the point contact between the semiconductor and the cat's whisker has a lower resistance to electrical flow in one direction than in the other. This difference in resistance accounts for the rectifying action of a crystal. Because it passes current predominantly in the direction of low resistance, the alternating voltage is converted to direct current.

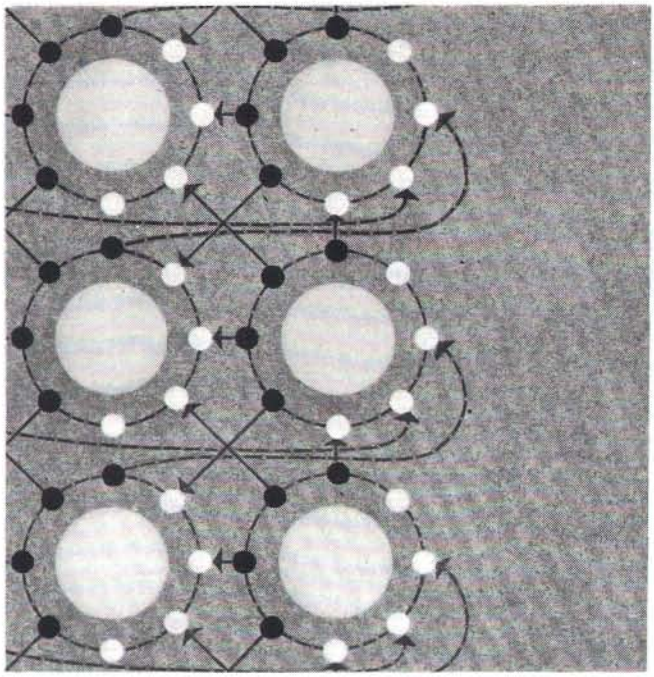
One would suppose that the respective resistances to current flow in one direction or the other would vary with the physical properties or resistance of the materials forming the contacts. But experiments showed that the properties of the metals made much less difference than the theory had predicted. Bardeen decided that something must take place at the surface of the crystal that the theory had not explained. Aided by previous work on a similar problem by William Shockley, director of the semiconductor research at Bell Labs, Bardeen undertook a theoretical study of the conditions at the surface of a semiconductor.

The result of this study was an important modification of the theory, which subsequent experiments were to prove correct. Bardeen reasoned that there were localized states on the surface of a semiconductor that differed from those in the interior. The number of such states, he said, was equal to the number of surface atoms. Like impurities in a crystal, these states produced holes capable of carrying current. These holes consisted of spaces on the exposed side of the atoms, which normally would be filled by electrons from adjacent atoms. This is an oversimplified picture of the theory, but it helps to make clear the essential concept: that the surface of a semiconductor is a better carrier of electricity than its interior. And Bardeen's theory satisfactorily accounted for the fact that the rectifying action of a crystal was independent of the particular metal used for the cat's whisker.

Shockley soon carried out an experiment that gave strong support to the theory. He reasoned that an externally applied electric field should increase the conductivity of a crystal by inducing electrons out of the bonds. He placed a sheet of germanium in an intense electric field. The increase in conductivity of the germanium turned out to be less than the old theory predicted. But the measurements fitted in well with Bardeen's new theory. They could be explained by the assumption, suggested by his theory, that the conductive layer of electrons or holes on the surface of the germanium acted as a shield against penetration of the ma-



ATOM OF BORON, also introduced into germanium, has one fewer outer electron. This "hole" is able to migrate through a crystal in much the same way as a real electron.



SURFACE ATOMS, one side of which does not adjoin other atoms, have unfilled holes. These make the surface of a crystal a conductor of tiny currents that pass across it.

terial by the electric field, just as metallic shields around parts of radio sets keep away stray electric fields.

Bardeen, Shockley and a colleague, W. H. Brattain, proceeded to further experiments and calculations, each new experiment resulting in refinements of the theory. They concluded that the superior conductivity of the surface layer of germanium in their experiments was accounted for chiefly by the presence of holes and that these holes were produced not only by impurities and surface states but also by the current passing through the crystal.

These studies, indicating a method of controlling the electrons or holes in a crystal, led Bardeen and Brattain to the invention of the transistor. The device consists of two fine tungsten wires of which the tips, only two thousandths of an inch apart, rest on a germanium crystal soldered in turn to a metal disk. All these elements are housed in a metal cylinder that is connected electrically to the metal disk and crystal, thus forming the ground terminal. The cat's whisker wires are connected to pins that can be plugged into a socket.

An electrical signal, modified by a small positive "bias" voltage to place it in the proper state for action on the crystal, is transmitted to the crystal by one of the cat's whiskers, which is called the emitter. The current releases holes in the surface layer of the crystal. The positive holes, flowing over the surface, are attracted to the second cat's whisker, which is biased negatively.

The first whisker controls the number of holes flowing to the second whisker, just as a vacuum tube grid controls the number of electrons flowing to the plate. The second whisker, called the collector, absorbs the current carried by the holes and passes on the signal, amplified 100 times.

The amplification is partly due to the fact that a change in the incoming current to the crystal produces a greater change in the outgoing current. Most of it derives, however, from the great difference in resistance between the input and output ends of the circuit. The behavior of the electrons or holes is controlled in the crystal by superposing variations on the positive and negative biased voltages applied to the emitter and collector. Thus, the transistor is essentially a triode form of the crystal diode detector used in radio.

Engineers at the Bell Laboratories have demonstrated that the transistor can be used as a voice amplifier, a television picture amplifier, a pulse amplifier and an oscillator. They have even produced a superheterodyne radio receiving set operating completely without vacuum tubes. Transistors were used in the set's amplifiers and in the local oscillator; conventional germanium crystal detectors served as mixer and detector, and selenium rectifiers were used in the power supply. This set performed as well as a conventional five-tube superheterodyne receiver. Since a transistor radio has no vacuum tubes to heat up, a program comes in at full strength as soon as the set is switched on.

This instant response of the transistor is especially useful in pulse-type communication systems and in electronic computers. The transistor will also have a special value in electronic equipment that must operate continuously even when there are power failures. Such equipment, which includes telephone repeaters, fire and burglar alarms and the like, is generally equipped with batteries for an emergency power supply. The small power requirements of the transistor will make it possible to use batteries for a prolonged period.

Yet all these applications are less important than the fundamental new knowledge that has been gained about the structure and energy states of solid matter and the electrical behavior of the surface atoms in a semiconductor. Basic study of these phenomena has been undertaken not only at Bell Labs but at Purdue University, the University of Pennsylvania and the Radiation Laboratory at the Massachusetts Institute of Technology. The holes in the crystal lattice of atoms obviously are a promising subject for further investigation.

Large-Scale Integration in Electronics

Success in the miniaturization of electronic circuits opened the way to vast improvements in electronic devices. Here the early stages of the process are described

by F. G. Heath

The most volatile technology of the present industrial age is that provided by electronics. Since the introduction of the transistor in 1948—which in its day seemed a marvel of compactness compared with the glass vacuum tube—the size of electronic devices has been reduced by a factor of 10 roughly every five years. This works out to a compression approaching 100,000 between 1948 and 1970. Ten years ago [1960]; when the term "microelectronics" first came into use, a chip of silicon a tenth of an inch square might hold 10 to 20 transistors, together with a few diodes, capacitors and resistors. Today such chips can contain well over 1,000 separate electronic components. The chip shown in the photomicrograph on the opposite page measures 0.11 by 0.14 inch and contains 1,244 transistors, 1,170 resistors and 71 diodes.

The technology that produces such high-density electronic circuits is called large-scale integration, or LSI. Although the term has no precise definition, it is usually reserved for integrated circuits that comprise 100 or more "gates," or individual circuit functions, laid down with a density of 50,000 to 100,000 components per square inch. If the upper value could be achieved throughout a cubic inch of material (which may be done in another decade or so), the density of electronic components would be about a fourth the density of nerve cells in the human brain.

A great part of the stimulus for miniaturizing electronic circuits came from ballistic missile programs. As the microtechnology was developed, however, it was speedily applied to commercial computers, with the result that the central processing unit in the current generation of machines is frequently smaller than the input and output devices that connect the processor to the outside world. It now seems inevitable that microelectronic circuits, including LSIs, will soon find their way into a variety of new applications whose impact on everyday life—in the home, in the office, in the school and on the highway—will be profound. For many of these potential uses, the main attraction of the miniature circuits will be not so much their microscopic size as their low cost per function and their high reliability. Because the integrated-circuit technology uses optical and photolithographic methods for creating transistors and other circuit components, costs are related directly to the number of steps involved and hardly at all to the pattern or density of the images produced. For comparable production runs, therefore, it should cost little more or no more to fabricate 100,000 transistors on a chip the size of a postage stamp than to fabricate 100 or 1,000.

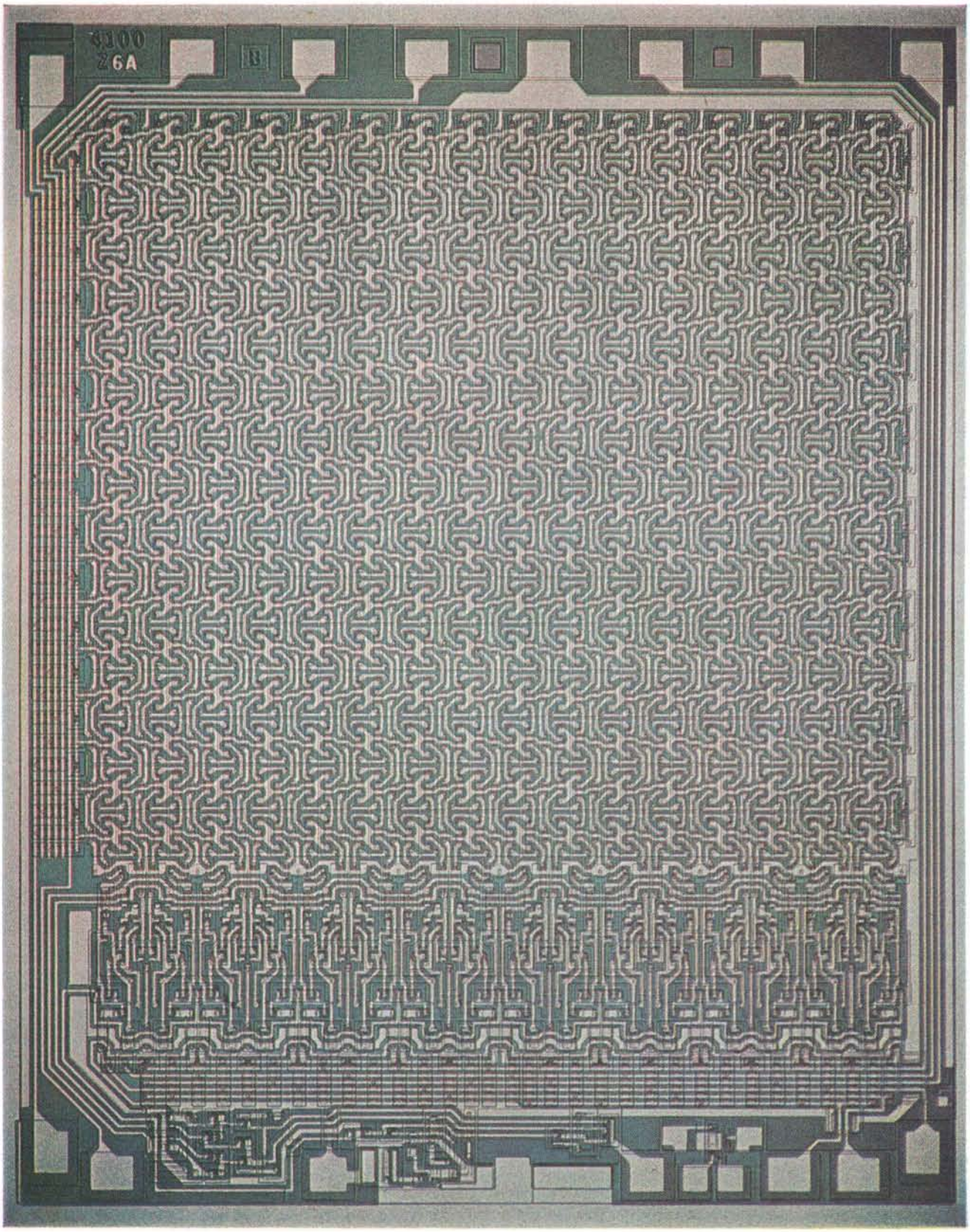
Until the advent of the transistor, each type of component in an electronic circuit was fashioned from one or more materials having the required electrical characteristics. For example, carbon was used for resistors, ceramics as a dielectric for capacitors, tungsten for the emitters in vacuum tubes and so on. These components, with characteristics defined by their composition and construction, were then used like building blocks in creating a circuit with specified characteristics and responses. Circuits were combined into systems, such as a radio transmitter, a radio receiver, a radar set or a computer.

From the earliest days, electronics has been a technology of complex interconnections. A small radar set can easily have as many connections as an oil refinery. The refinery, however, may cost millions of dollars and require a maintenance crew around the clock. The radar set, on the other hand, may cost only a few thousand dollars and is expected to work reliably for weeks on end. When it does need repair, the job will be done by a repairman who may not have seen it for many months. The complexity and inherent reliability of electronics is taken largely for granted.

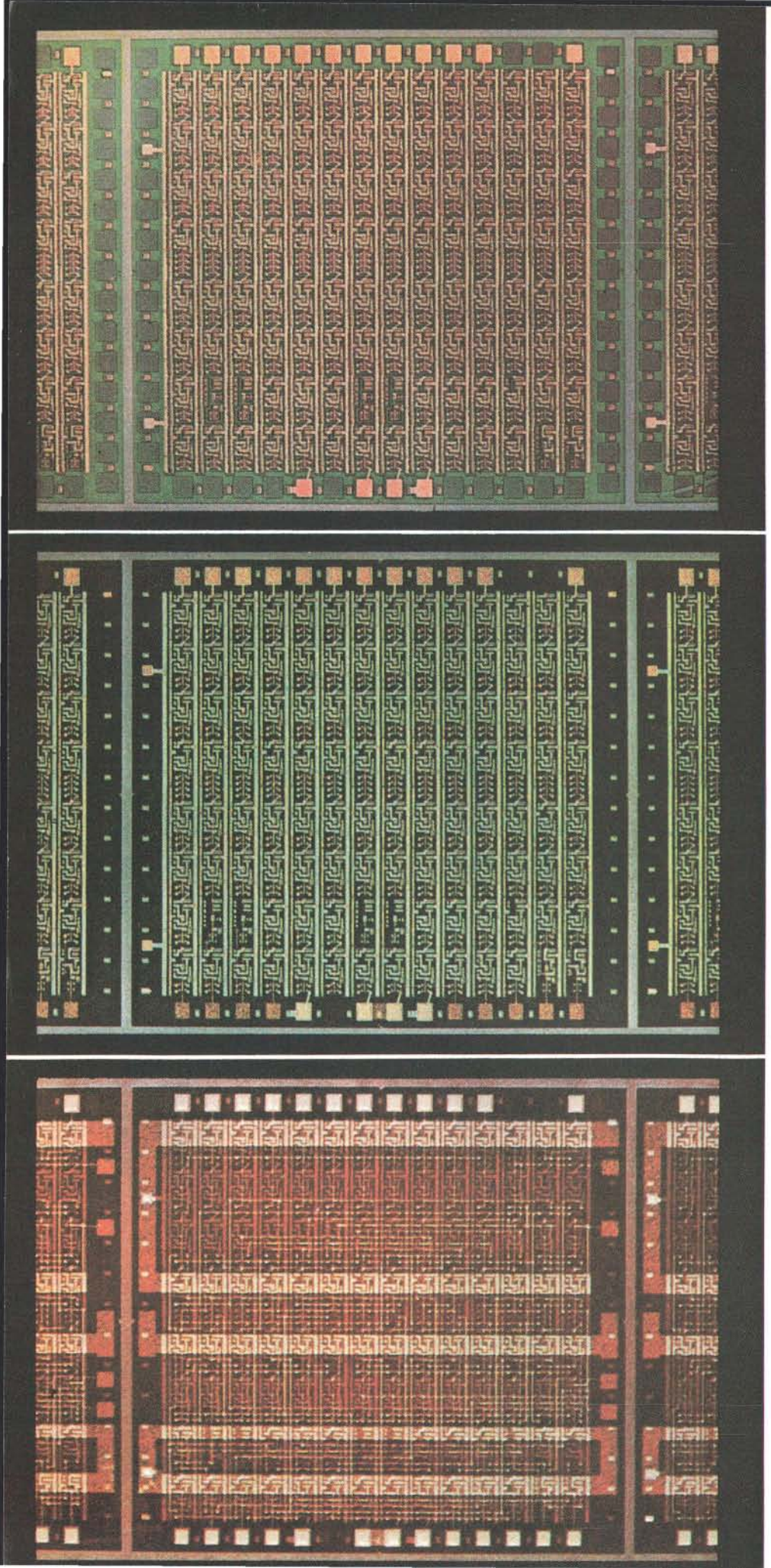
The great advantage of the transistor, an advantage scarcely appreciated at first, was that it enabled one to do away with the separate materials—car-

F. G. HEATH was professor of digital processes at the Institute of Science and Technology of the University of Manchester and scientific adviser to International Computers Ltd. when he wrote this article in 1970. After completing his education at Manchester, he worked in the electronics industry designing radio tubes and photoelectric multipliers. Joining Ferranti Ltd., he worked increasingly in the area of digital electronics up to the time the firm produced the prototype Argus process-control computer in 1957. "When this machine was demonstrated to the Duke of Edinburgh," he said, "I as project leader had a most important job: to see that under no circumstances hot hydraulic oil squirted on the royal visitor." On his return to the University of Manchester, Heath served as lecturer in electrical engineering at the Institute of Science and Technology and as chief engineer at the Manchester Engineering Facilities of International Computers Ltd.

Reprinted from the February 1970 issue



COMPUTER MEMORY CIRCUIT, incorporating 1,244 transistors, 1,170 resistors and 71 diodes, is an example of large-scale integration (LSI). The term is usually reserved for integrated circuits whose density is 50,000 or more devices per square inch. This unit, manufactured by Fairchild Semiconductor, has a density exceeding 160,000 devices per square inch. Its actual size is 0.11 by 0.14 inch, roughly the area occupied by the word "the" in a typeface somewhat smaller than this one. The circuit can store 256 bits of information and, with similar units, will provide the high-speed memory for the advanced Illiac 4 computer now [1970] under development at the University of Illinois. It provides faster access than conventional high-speed memory units, which consist of tiny rings of ceramic threaded on wires.



bon, ceramics, tungsten and so on—traditionally used in fabricating components. At the same time, the transistor raised the ceiling that sheer complexity of interconnections was beginning to place on system design.

At first the transistor did little to alter the requirement for connecting individual components, but the technique of its construction did. The transistor was the first electronic component in which materials with different electrical characteristics were not interconnected but were physically joined in one structure. The preferred material for making the transistor soon came to be silicon; it was produced in the form of a single crystal resembling a sausage, which was sliced into thin round wafers. By suitable masking and "doping" techniques, which selectively altered the electrical behavior of small regions, several score transistors could be created on each wafer. The individual transistors were then cut apart and sealed in a package about the size of a pencil eraser. The problem of soldering individual transistors and diodes into a circuit remained.

As a first step toward simplifying system design and reducing the number of interconnections, computer engineers began to develop a series of standard circuit modules, each of which per-

COMPUTER LOGIC CIRCUIT is shown in three stages of manufacture. Each circuit contains 112 identical cells, or logic "gates," made up of three transistors and four resistors. The complete circuit, which measures 0.133 by 0.142 inch, thus contains 784 devices, equivalent to an average density of about 40,000 devices per square inch. The 112-gate circuit, which fills most of each picture, is only one of about 100 created on a single chip of silicon. The three pictures show three stages of metallization. The first metallization (*top*) connects the transistors and resistors in the individual gates; the second metallization (*middle*) interconnects the gates, and the third (*bottom*) provides the leads for the power supply. The interconnections can be patterned in many ways to provide custom circuits. The circuit is manufactured by Motorola Semiconductor Products, Inc. Like the circuit on the preceding page, it employs the "bipolar" technology. The newer metal oxide-semiconductor (MOS) method provides a still higher density of circuit elements. The two techniques are compared in the illustrations on pages 160 and 161.

formed a specific function, and used them as logical building blocks for creating their systems. The transistor, being much smaller than a vacuum tube, could readily be assembled with resistors and capacitors of about the same size on a small plastic board. These modular circuit boards, typically the size of a playing card, could then be plugged together as needed.

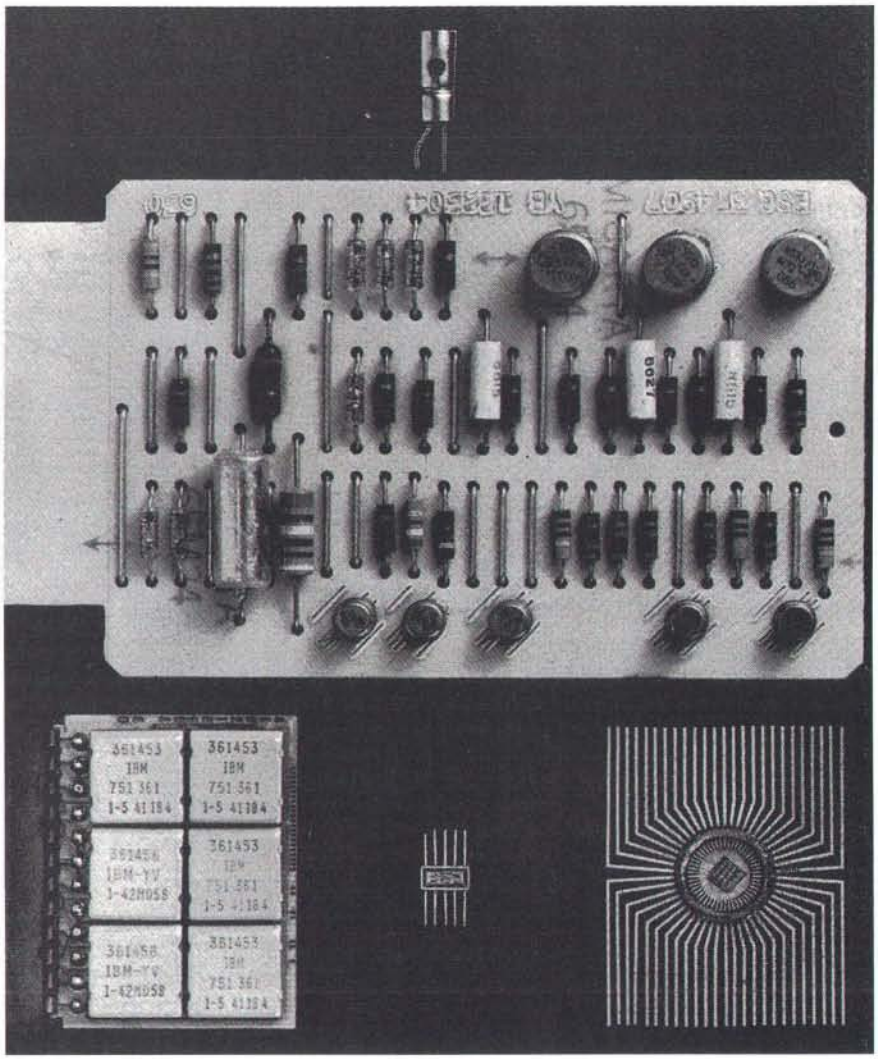
With the growing complexity of systems, however, the interconnection problem again asserted itself. In small systems, engineers had been able to position individual components quite freely, worrying only about electromagnetic interference from such components as transformers. In the much larger new systems, the fabrication of complex wiring networks became very costly. What was even more important, the increased speeds of operation possible with transistors were in fact outstripping the speed with which signals could travel along the interconnecting wires. It was rather like trying to drive a 150-mile-an-hour sports car behind a long queue of traffic traveling at 40 miles an hour; the shorter the queue, the easier it is to overtake the cars ahead and move to the front of the queue. It is much the same with electronics. By keeping the interconnecting wires as short as possible, one can step up the speed of operation.

As transistor technology developed and the switching speed of the circuits became faster, it was increasingly important to decrease the size of components and the length of interconnections. The physical limit of finding room for connections in an ever decreasing area was approaching rapidly. This limitation, coupled with the increasing complexity of system design, made the search for a new technology imperative.

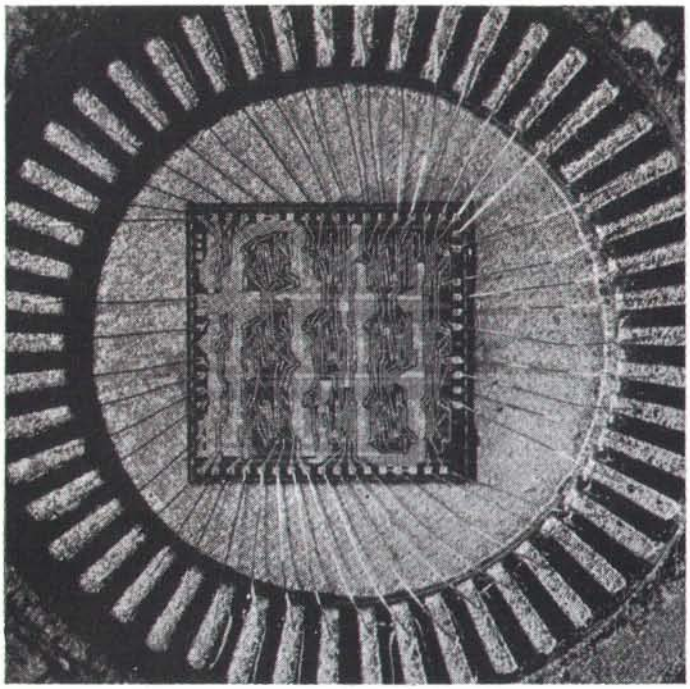
The technology that resulted was microelectronics, embodied in the integrated circuit. Beginning in the mid-1950s, engineers in two American companies, Texas Instruments and Fairchild Semiconductor, saw the possibility of producing as part of a single chip of silicon not only transistors and diodes but also resistors and capacitors and of joining them into a complete circuit. The special properties needed for the various circuit elements were achieved by selectively diffusing traces of impurities into the silicon or oxidizing it to silicon dioxide. The principles of photolithography were used to expose selected regions of the silicon to diffusion while protecting other regions [*see top illustration on pages 160 and 161*].

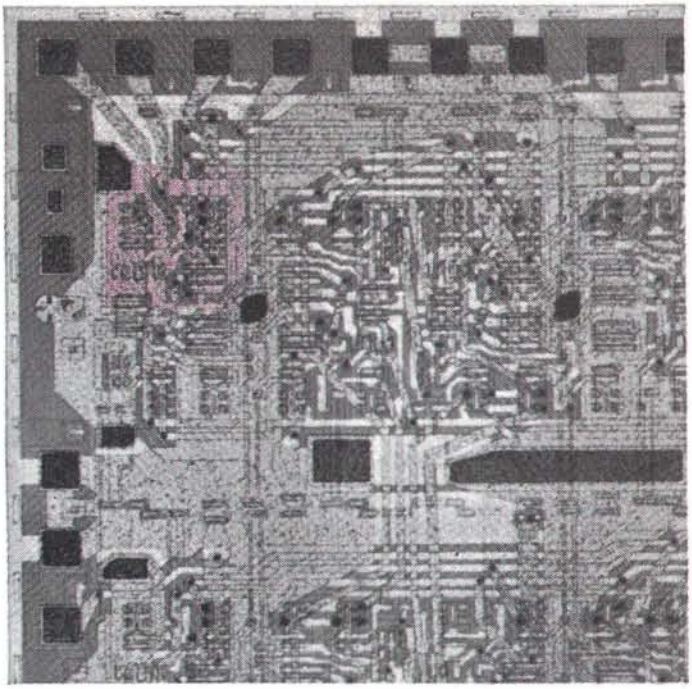
At first Texas Instruments used fine wires for bonding the various elements into a functional circuit. Fairchild achieved the same result more simply by evaporating a thin film of aluminum over the circuit elements and etching it selectively to leave a two-dimensional network. The Fairchild technique produced what became known as planar integrated circuits.

The standard integrated circuit incorporates between 100 and 500 identical logic circuits on a thin slice of silicon in an area roughly an inch square. These circuits are cut up into individu-



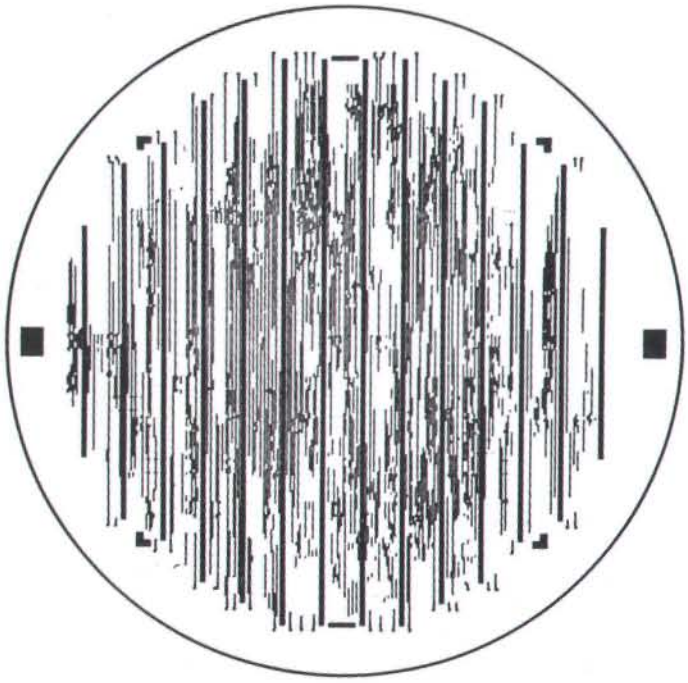
EVOLUTION OF MICROELECTRONICS began with the introduction of the transistor in 1948. A 1952 transistor is shown at the top. (All the objects are depicted at actual size.) The next step (*directly below*) was to assemble several transistors, diodes, resistors and capacitors on a circuit board that could be plugged into a computer or other kind of system. The example shown here is a circuit board of the type used in the 7,000-series of IBM computers in the early 1960s. It incorporates 45 electronic devices. The IBM 360-series of computers employs a technology called solid state logic (*bottom left*); it provides 40-odd devices in six tiny cans. The late 1950s saw the development of the first integrated circuits in which transistors, diodes and resistors were all formed on a single chip of silicon and linked into a circuit. The example in the middle of the bottom row was made in 1965 by Texas Instruments; it contains 91 transistors and resistors. The LSI example at the bottom right, made by Fairchild, contains 864 devices. This unit, in various magnifications, reappears at the top of the next two pages.

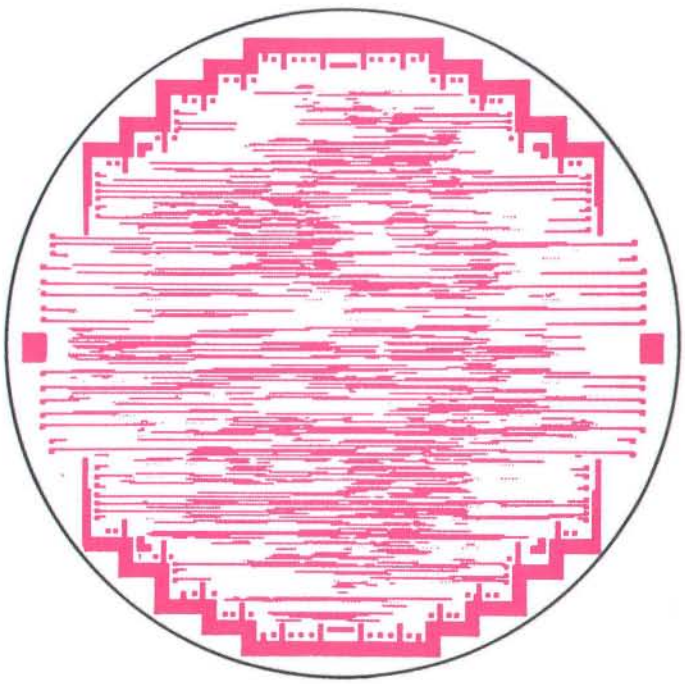




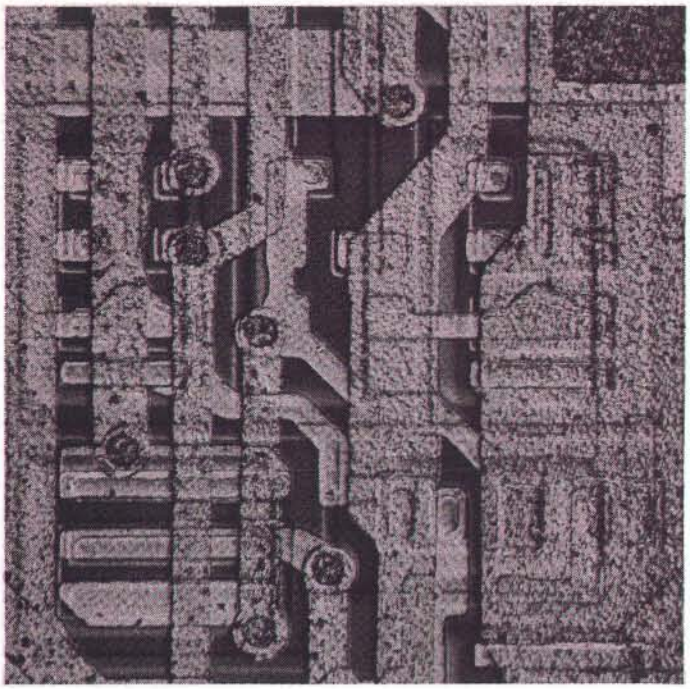
SEQUENCE OF MAGNIFICATIONS reveals the fine structure of circuits made by the large-scale-integration technique. The Fairchild circuit depicted in the illustration on the preceding page is shown here magnified 10, 50 and 250 diameters. The

al units. Each unit is sealed in a package containing anywhere from half a dozen to two dozen minute connections needed to bring the circuit in contact with the outside world, which is to say in contact with the scores, hundreds or thousands of similar building blocks that are required for a complete signal-processing system.





DISCRETIONARY WIRING is employed by Texas Instruments in some of its LSI circuits. The one shown here contains 250 gates, but as many as 4,000 could be placed on a single silicon wafer 1.5 inches in diameter. A 4,000-gate array would require at least 12,000 transistors. In this approach, which has low cost as a goal, the devices are individually tested and defective units are recorded in a computer memory. The computer then figures out an efficient wiring plan to link as



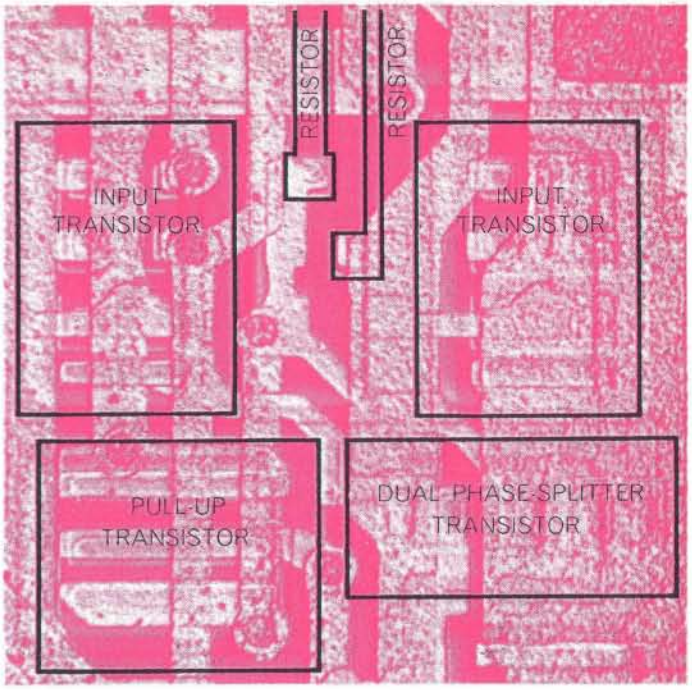
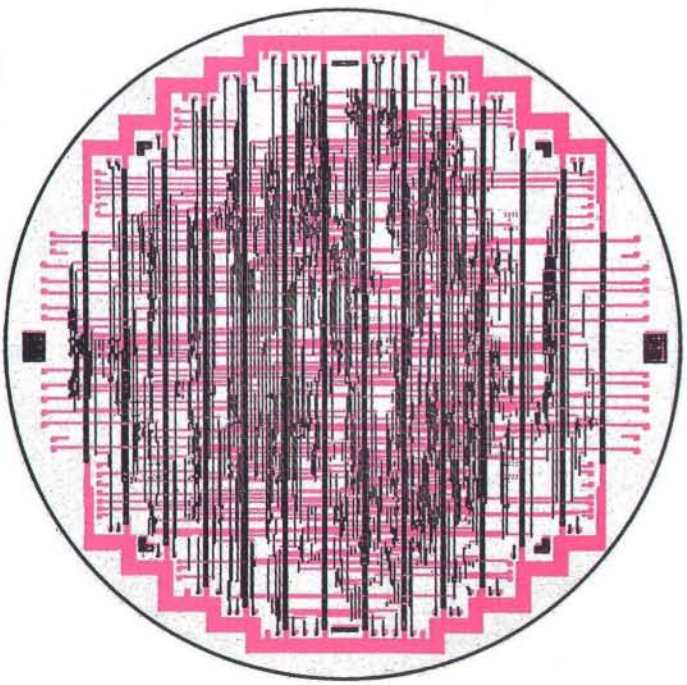
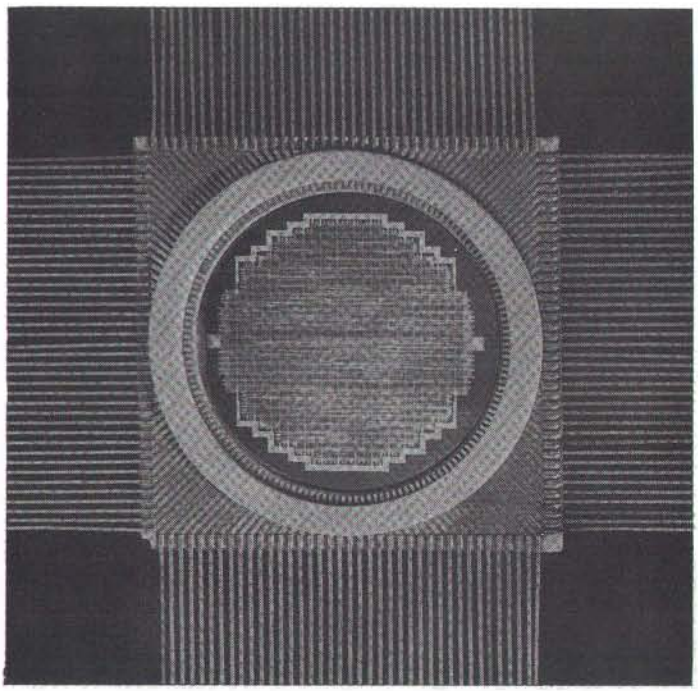


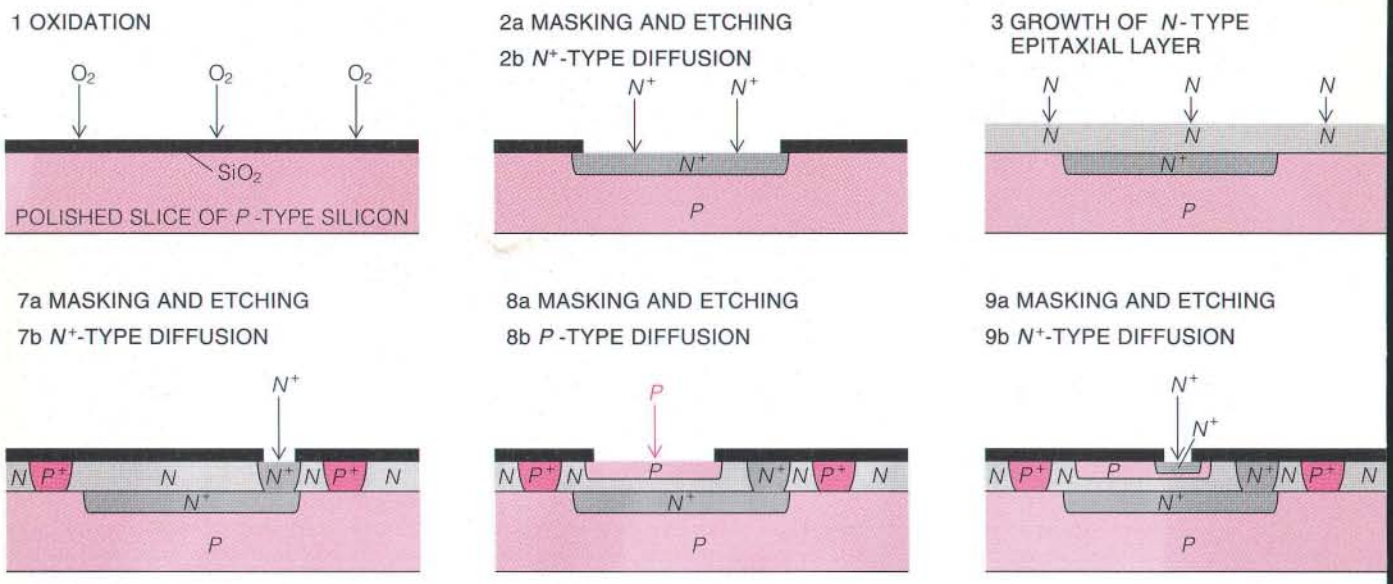
diagram at the far right shows the location of individual transistors and resistors in the 250-diameter enlargement. The complete circuit contains 864 devices, packed with a density of more than 40,000 per square inch.

The concept of using circuit building blocks as "logic" units in a computer can be traced back to George Boole, who demonstrated in 1854 that all mathematical and logical processes can be synthesized by a binary system using only three operators: "and," "or" and "not." (Boole would probably be appalled by this simplification as well





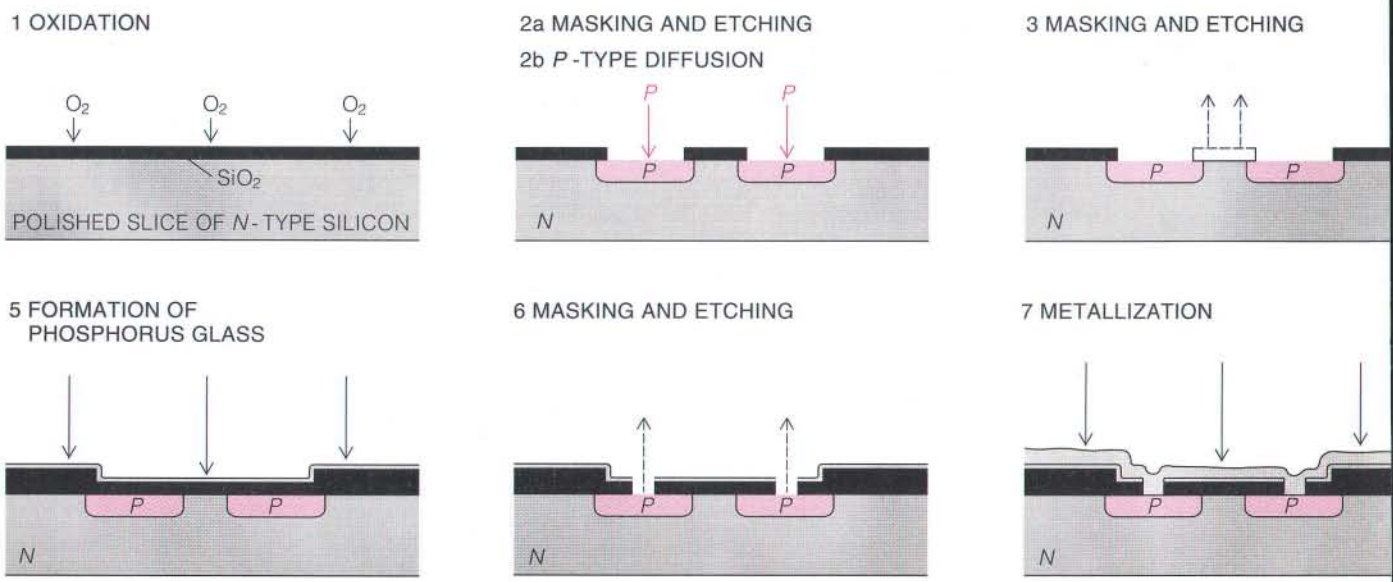
many good devices as needed to produce a circuit of a particular design. The process begins when a cathode-ray tube generates vertical and horizontal wiring patterns separately, as shown in the first two parts of the illustration; together those patterns provide the complete wiring scheme, which appears in the third pattern. The finished large-scale-integrated circuit package as it was manufactured by Texas Instruments is shown at actual size at the far right.



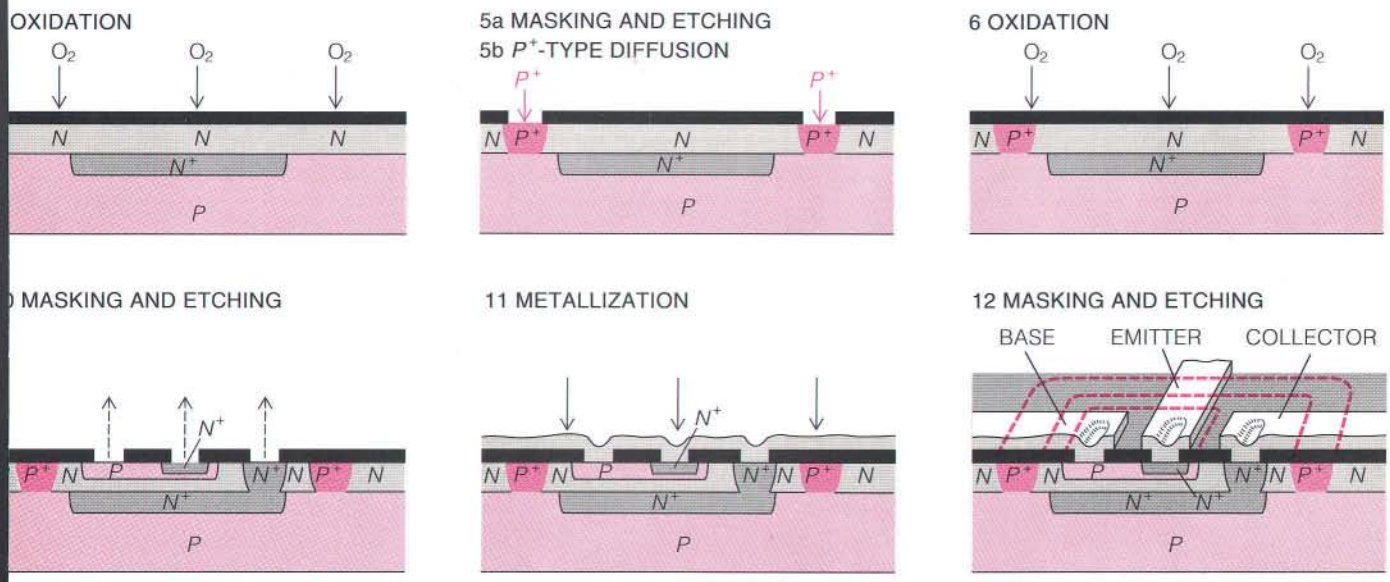
CONVENTIONAL INTEGRATED CIRCUITS are made by the bipolar technique, in which transistors are formed from *n* (negative) and *p* (positive) regions. In *n* regions free electrons are the natural carriers of electricity; in *p* regions "holes" with positive charge are the natural carriers. The process begins with a polished wafer of silicon containing impurities of the *p* type. Oxidation produces a thin layer of silicon dioxide (*1*). The next step, photoetching (*2a*), is fundamental to the entire process of circuit fabrication. Photoetching creates the high-resolution patterns essential for making microelectronic circuits. As in photolithography, the wafer is coated with a resistant material that hardens on exposure to light. An extremely accurate mask, produced by precision optics, is placed between the light source and the wafer. In areas not exposed to light, the resistant material, called the resist, is dissolved and the underlying silicon dioxide is removed by

as fascinated by what computer designers have achieved with his concepts.) The central processor of a computer typically employs several thousand of these logic operators, realized by the circuits termed gates. The general theory of the functioning of more than, say, six interconnected gates is quite beyond human comprehension.

For lack of such a theory, computer engineers have to work with simplifying concepts that enable them to un-



NEWER TYPE OF INTEGRATED CIRCUIT employs the metal oxide–semiconductor technology. It can create circuits with five to 10 times as many transistors per unit of area as the bipolar technique. The process starts with a wafer of *n*-type silicon, which is oxidized (*1*). The openings that will provide the "source" and "drain" of a field-effect transistor are photoetched (*2a*) and exposed to *p*-type diffusion (*2b*). The gate region is removed (*3*), and a new oxide layer is created over the entire surface (*4*). For purposes of stabilization, a thin layer of phosphorus glass is formed (*5*), and windows are photoetched to expose the source and drain (*6*). The surface is aluminized (*7*) and given a final photoetching (*8*). Transistors



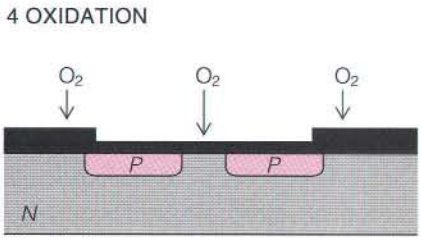
etching. The wafer is then placed in a controlled atmosphere so that n^+-type impurities can be diffused into the formerly p-type silicon (*2b*). Next the silicon dioxide layer is removed from the entire wafer and a mild n-type epitaxial layer of silicon is grown on the surface (*3*). "Epitaxial" signifies that the underlying crystal structure is continued without interruption. The surface is again oxidized (*4*), etched (*5a*) and exposed to a p^+-type diffusion (*5b*). This is followed by oxidation (*6*), etching (*7a*) and an n^+-type diffusion (*7b*). After a second oxidation process, which is not shown, the surface is again etched (*8a*) and exposed to a p-type diffusion (*8b*). After another oxidation and etching (*9a*) there is a final n^+-type diffusion (*9b*). Another etching (*10*) produces the openings needed for a layer of aluminum (*11*) to make contact with the p and n^+ regions of a transistor. A final etch leaves a network of leads on the surface (*12*).

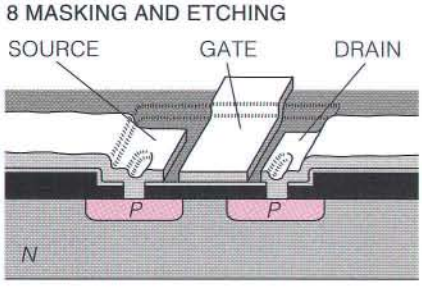
derstand and manipulate large chunks of logic. One method is to organize computer operations around "registers" in which binary digits represent a particular quantity or number. These registers and the connections between them specify the data flow; they are organized in regular structures. The extreme example is provided by the computer memory, which consists of vertical stacks of identical registers. Data are moved from one register to another by a series of timing signals. The sum total of such signals defines the computer's microprogram. It is largely the microprogram, with its near random properties, that complicates the interconnection pattern in a large computer. The more modular the physical construction of a computer can be made, the easier it is to create a satisfactory system. The use of standard integrated circuits eased the circuit-design and interconnection problem at the lower level, but at the higher level serious problems remained.

If, for example, one wanted to design a central processor with 1,000 logic elements, one could begin by selecting 200 integrated-circuit packages, each containing an average of five gates of the kind desired. Tying the 200 packages together, however, might easily require 6,000 connections. Since each connection must be kept as short as possible, the solution depends on two things: first, a good miniature wiring method and, second, computer programs that can transform the designer's logic into a good layout with simple interconnections. The conventional two-sided printed circuit board was incapable of doing this job.

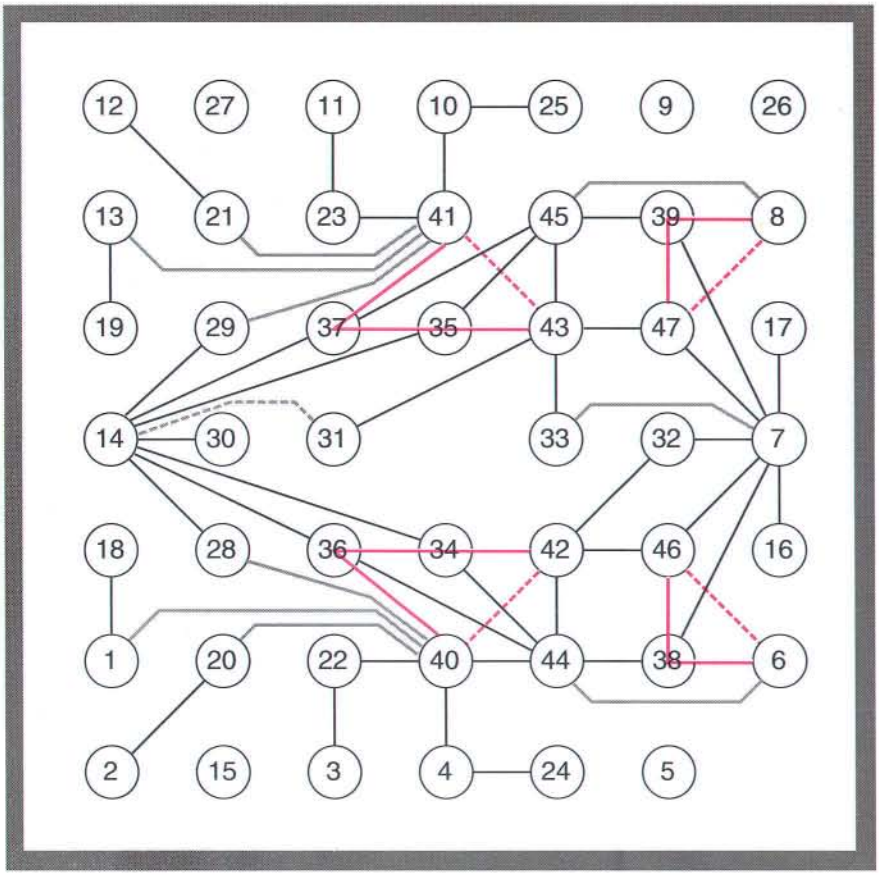
A practical solution to the problem was found by designing multilayer printed circuit boards about a foot square on which 100 to 200 integrated circuits could be mounted. The board is a laminate of double-sided printed circuits, each fabricated to an accuracy of about a thousandth of an inch. Typically 12 layers of wiring crisscross through an insulation of reinforced plastic. This precise and rigid three-dimensional structure replaces the bird's nest of wires found inside earlier computers. Since it is not humanly possible to lay out 6,000 connections in an area a foot square and 12 layers deep, computer programs had to be developed to do the job. This involved an immense programming task requiring hundreds of thousands of words of instructions. The end result was that a very large central processor could be assembled from only about 60 multilayer boards, linked together by relatively few hand-wired connections. The task of testing such high-density electronic structures was solved with the help of automatic testing machines.

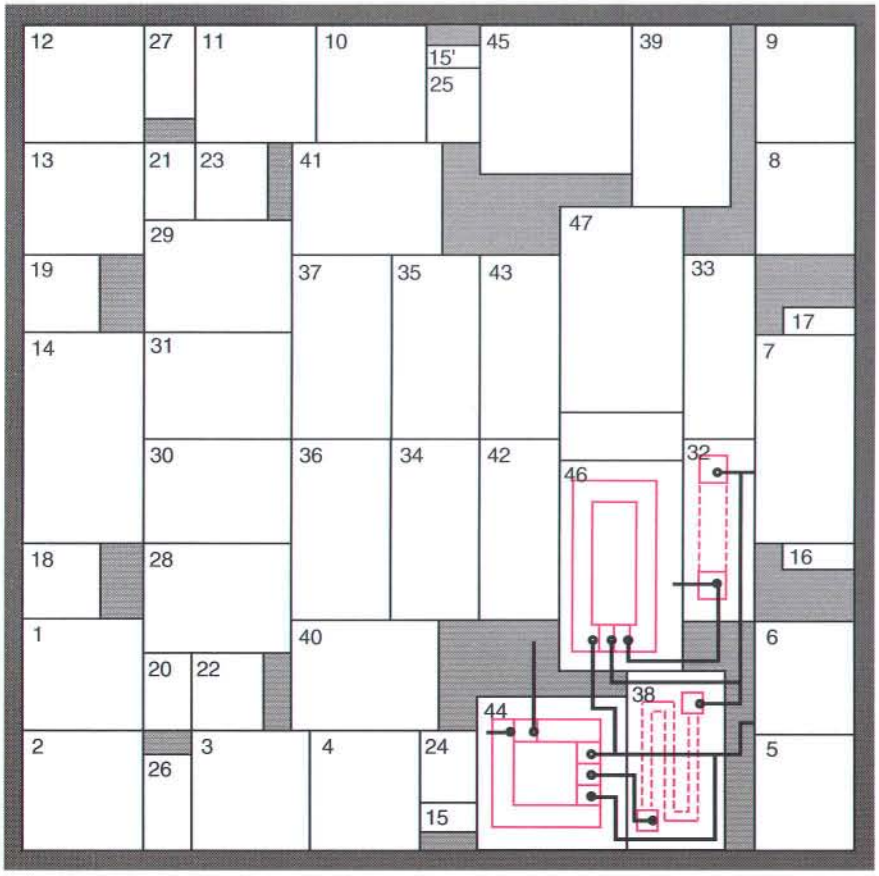
One can see that no great leap of the





produced by the MOS technique, unlike transistors manufactured with the bipolar technique depicted at the top of these two pages, have electrical characteristics quite similar to those of old-fashioned vacuum tubes.





imagination was required for the next step: large-scale integration. The makers of integrated circuits had learned how to produce chips about an inch square that held several hundred logic circuits, which were then cut apart and packaged individually. Why not fabricate a chip in which the several hundred circuits were interconnected and sold as a single large package?

One obstacle, and a major one, was that a multicircuit chip would have to incorporate logic gates of several different types and all would have to function perfectly. When the manufacturer had put several hundred identical gates on one chip, each one could be tested individually and any defective ones discarded. A 60 or 70 percent yield was considered good.

One way around this obstacle to large-scale integration was not long in coming: lay down more gates than needed, identify the bad ones by test-

COMPUTER-AIDED DESIGN has become essential for arranging the devices and interconnections in LSI circuits. At the University of Manchester Institute of Science and Technology, the author and his associates have worked out a series of computer programs that make the design of interconnections almost automatic. In the first step of the procedure, the circuit designer prepares a logic diagram of his needs and specifies the individual circuit elements required. This information is punched on paper tape and fed into a computer. A program called PLACE examines the interconnections required and attempts to locate elements to give the simplest connection pattern. In the top diagram at the left the elements have been moved from their original arbitrary locations, represented by the numbers in the circles. (Thus, element No. 1 had been the first element in the first row, and so on.) At this stage the connections are shown as straight black lines. For simplicity, not all connections are shown. Program CROSSOVERS lists the places where a connection crosses an element, and CROSSBOW bends the connections to one side (*gray lines with angles*). Program DUCK identifies certain crossings (*broken lines in color*) that can be avoided by routing the lines so as to pass under resistors (*solid lines in color*). After several more programs PACK examines the shape and size of the various elements and fits them into the smallest possible area, as shown in the bottom diagram. Gray regions denote unfilled spaces; the outlines of typical elements are in color. Finally, a program called LEE (named for its developer) fits in the actual wiring (*heavy black lines*).

ing and interconnect only the good ones by multilayer metallization. Again computers were called on to supervise the testing and to lay out custom interconnections for each chip. It also became a simple matter to interconnect the gates in a chip so as to meet the special requirements of different customers. The result was LSI with discretionary wiring [*see bottom illustration on pages 158 and 159*]. So far manufacturers have achieved three layers of metallization. The more layers of connections, the more the components can be packed together, thus improving the yield and reducing the cost. Recently manufacturers have succeeded in producing LSI units in which all the gates are usable.

Faulty circuits, which reduce the yield, are caused by a point fault on the semiconductor crystal; such faults are randomly distributed over the surface of the silicon chip. A newer type of transistor element—the field-effect transistor—has come into use because it has better properties, occupies less area on the chip and therefore is less likely to be spoiled by a point fault. These newer kinds of transistor are produced by the metal oxide–semiconductor (MOS) technique [*see bottom illustration on pages 160 and 161*]. The MOS technique can provide circuits some five to 10 times more complex than the "bipolar" technique used for making conventional transistors and still achieve the same yield. There are, however, subtle operating differences in the transistors produced by the two techniques, with the result that one or the other may be preferred, depending on the particular circuit application. Recently ways have been found to mix the two techniques, so that circuit versatility is enhanced.

The production of large-scale integrated circuits with computer logic gates is still modest in volume. A number of companies have directed their initial LSI effort toward producing computer memory arrays, which have a highly regular structure and potentially a very large market. In most memory arrays under development, all the gates must operate, so that very high standards of production are required to attain satisfactory yields. By 1972, if not sooner, it should be economic to replace the standard magnetic-core memory of the computer (the type of memory in which tiny magnetic rings of ceramic are threaded on a matrix of thin wires) with a series of LSI memory modules. Each LSI memory module will use chips with about 1,000 gates and will hold 256 words, each 32 "bits" long. Between 25 and 50 of these modules, interconnected by some suitable microwiring method, should be able to fit in a cube no more than four inches on a side. This is an achievement that would have been inconceivable even five years ago.

The development of LSI discretionary-circuit modules and LSI memory modules offers an exciting prospect for the system designer. It means that the central processor for a high-capacity computer could be built from perhaps 100 LSI chips, making such a computer readily portable. It will be difficult, however, for the LSI-chip manufacturer to quote low prices if all 100 chips have to be different. To minimize the variety of chips needed, system designers are investigating new computer system structures. For example, it should be possible to design computers consisting of blocks of identical LSI chips in which arithmetical operations are performed in parallel rather than sequentially, as they are at present. Such parallel systems would be ideal for time-sharing applications, where large numbers of users want simultaneous access to the same computer. Because each LSI building block is virtually an independent processor in its own right, one can readily imagine LSI chips joined together in various ways to provide computers tailored to each customer's special needs. The end result of LSI technology will be to increase computer processing speeds, to reduce the amount of basic programming required, to increase the capability of small processors built into terminal devices (thus reducing the load on the main computer) and to reduce costs all around.

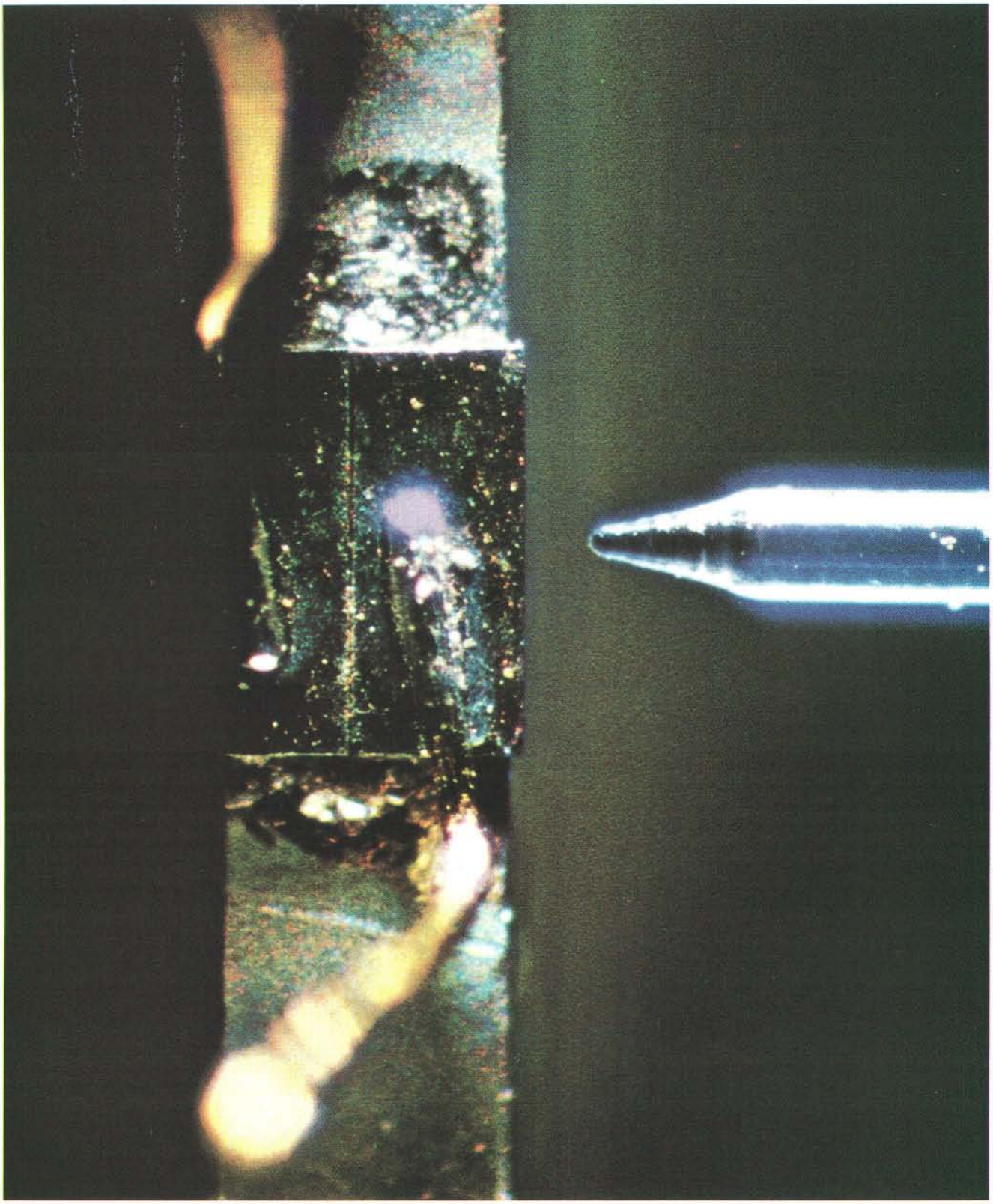
To give the reader some idea of how an LSI chip is designed, let me describe briefly a computer-aided design program we have developed at the University of Manchester Institute of Science and Technology, with the support of the International Telephone and Telegraph Corporation. The designer first prepares a diagram of the logic operations required in the finished LSI. This is punched on a paper tape and fed into a computer, to be used in conjunction with a library of standard circuit elements. A program called PLACE examines the interconnections needed and locates the elements to provide the simplest connection pattern. At this step the elements are connected by straight lines.

Next a program called CROSSOVERS notes every place where straight connections cross over an element. CROSSBOW moves all these connections so that they no longer cross the element below. TWIST tries out each element in all orientations so that the wires out of each element are correctly positioned. DUCK takes any crossings that remain, identifies where a resistor is involved and routes the connection to pass below the resistor. The crossings that cannot be handled in this fashion are examined further by SEARCH, which seeks other ways to avoid a crossing. The final residue of crossovers is handled by CHEAT, which creates a minimum number of low-value resistors that can be placed under the surface and used as crossing devices.

A program called PACK examines the dimensions of each element and fits all the elements into the smallest possible area without overlap. If there is not enough room between the elements for the connecting wires, STRETCH expands the layout as needed. Finally, LEE (an algorithm developed by Chi-Yuan Lee of the Bell Telephone Laboratories) fits in the actual wiring, using the rules of the particular LSI technology being employed. Programs of this kind will undoubtedly be capable of increasingly complex tasks—the first gropings, perhaps, toward artificial intelligence. Stimulated by these exercises in automated circuit design, we are now examining techniques for organizing an entire factory on a logical basis, not necessarily with the goal of automating the operation but to simplify both human and machine functions in a manufacturing complex.

It must not be forgotten that semiconductor techniques can be applied not only to circuit making but also to the production of such things as photoelectric cells and light sources. If 250,000 components can be put into a volume the size of a package of cigarettes, one could hope to duplicate an insect's eye by interconnecting thousands of tiny photocells with the LSI technique. A similar development could produce a flat, solid state television picture screen.

Such is the potential for many new developments made possible by LSI: wristwatch television, robot toys a few inches high, a computer terminal for the home, electronically guided automobiles. One can predict that new markets of this kind will lead to a drastic reduction in the cost of LSI circuitry. The decade of the 1970s should be a boom time for microelectronics.



CLEAVED-COUPLED-CAVITY LASER developed by the author and his colleagues at AT&T Bell Laboratories was photographed under a microscope there. At the right is an optical fiber roughly 100 microns in diameter. Its end is lensed, that is, shaped to capture laser light. The tip of the fiber confronts the laser (the rectangular object at the left). The thin vertical line on its surface is a gap between two half lasers; the cleaved-coupled-cavity laser (called for that reason the C^3 laser) consists of two lasers with a space between them. The C^3 laser emits light at the invisible wavelength of 1.55 microns. At that wavelength a fiber made of silica glass is most nearly transparent to light.

The C^3 Laser

The alignment of two conventional semiconductor lasers yields a beam of almost perfect purity that enables communications systems to send signals at rates as great as billions of bits, or binary digits, per second

by W. T. Tsang

When the Indians of North America transmitted messages by means of smoke signals, they were exploiting concepts at the heart of modern optical communication. The intermittent puffs of smoke they released from a mountaintop were digital signals. Indeed, the signals were binary: they encoded information in the form of puffs of smoke or the absence of puffs of smoke. Light was the information carrier; air was the transmission medium; the human eye was the photodetector. The duplication of the signal at a second mountaintop for transmission to a third served as signal reamplification.

Today the digital signals are pulses of light produced by a semiconductor laser; the transmission medium is fiber optics. Indeed, it is the simultaneous achievement of reliable semiconductor lasers and low-loss optical fibers that will enable communications systems to carry the increases in information traffic expected for the balance of the century. The superiority of an optical system over an electrical one is measured by criteria including information-carrying capacity (four orders of magnitude greater for an optical system), energy loss in signal transmission (two orders of magnitude lower) and error rate (one order of magnitude lower). In the U.S. the American Telephone and Telegraph Company had some 20 million circuit miles of optical communications lines in operation or being installed at the end of last year [1983]. An undersea optical line between North America and Europe is planned for service beginning in 1988. [It began in 1989.]

Nevertheless, no conventional semiconductor laser manufactured today is ideal for optical communication. For one, the lasers cannot exploit the full carrying capacity of the optical fibers. The available semiconductor lasers—contrary to the popular understanding—do not emit light at a single wavelength. They emit instead at a family of wavelengths, and different wavelengths are transmitted at different velocities in optical fibers. The result can be the blurring of a signal.

Here I shall describe a new laser that my colleagues and I have developed at AT&T Bell Laboratories. We call it the cleaved-coupled-cavity, or C^3 laser. In essence it is no more than the alignment of two conventional semiconductor lasers that can interact optically but are electrically independent. By means of this alignment the laser becomes capable of purifying its own output, so that what emerges is a beam of electromagnetic radiation at a single wavelength. In addition, the laser becomes tunable, so that the output can be switched with great rapidity from one wavelength to another. The C^3 laser promises great improvements in the information-carrying capacity of optical communications systems. Moreover, the laser's tunability will open new applications beyond those in optical communications. For example, the laser can serve as an optical-electronic logic element processing information at a rate of gigabits (10^9 operations) per second.

Since the central technologies of optical communications—the laser, or signal producer, and the optical fiber, or signal transmission medium—are intimately related, an explanation of the usefulness of the C^3 laser begins not with the laser but with the fiber.

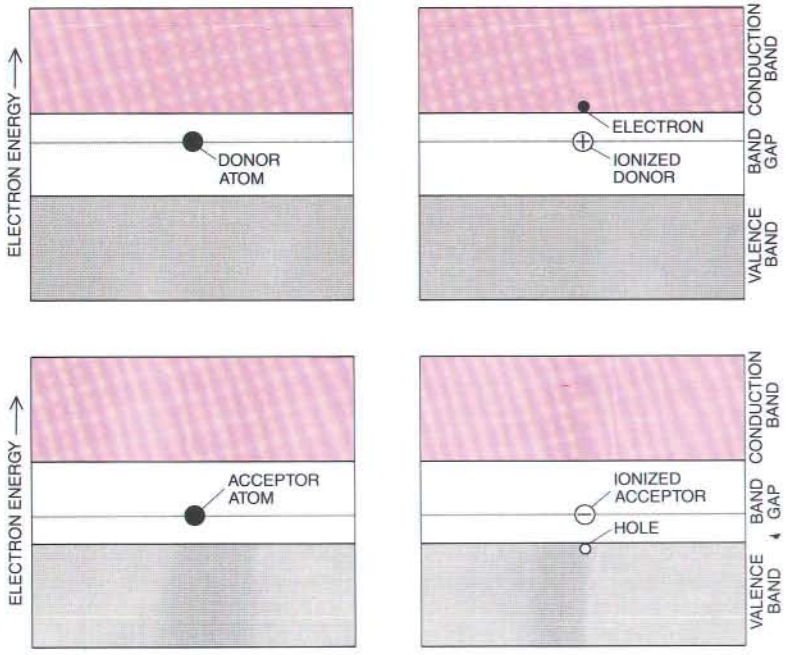
An optical fiber carries light along its length because light cannot escape from it. The light cannot escape because the refractive index of the core of the fiber is higher than the refractive index of the cladding that surrounds the core. (The refractive index is simply the ratio between the speed of light in a vacuum and its speed in a material.) Because of the gradient in refractive index between the core and the cladding, the light is trapped; it follows a zigzag path down the length of the fiber.

The number of possible paths (or modes of propagation) in any given fiber is determined by the diameter of the core, by the gradient in refractive index from the core to the cladding and by the wavelength of the light. When the diameter or the gradient is large, many paths are available; such fibers are multimode. When the diameter or the gradient is reduced, fewer modes of propagation can be accommodated. Eventually only a single, axial mode is allowed.

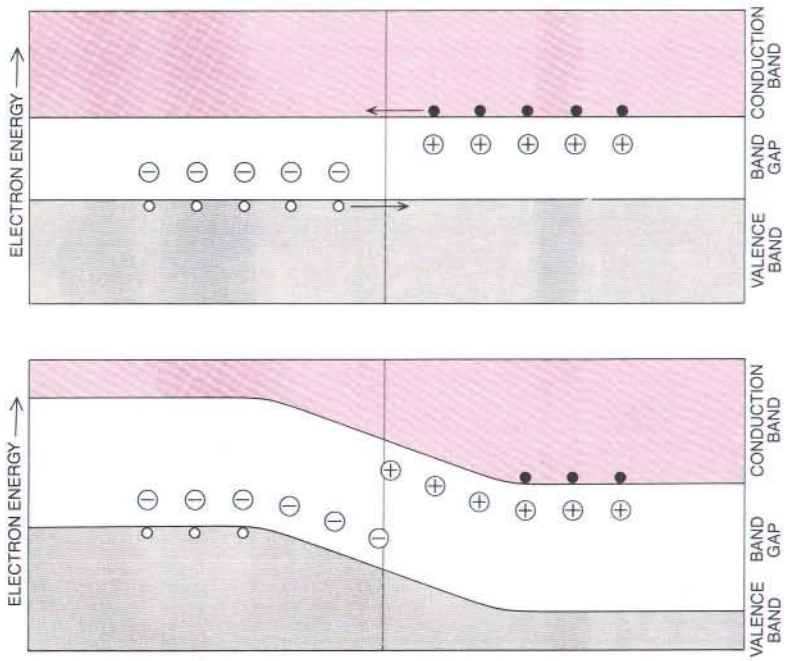
In a multimode fiber a pulse of monochromatic light travels by all the modes, that is, by all the possible zigzag paths. Each path has a different length, and so each mode entails a different transmission time. As a result, the pulse broadens, or smears out, over time. Since information is encoded by patterns of pulses, the smearing limits the rate of information transmission (expressed as bits, or binary digits, per second) and also the spacing of repeaters (signal reamplifiers). The problem can be avoided by employing a single-mode fiber.

W. T. TSANG is head of the Semiconductor Electronics Research Department at AT&T Bell Laboratories in Murray Hill, N.J. A native of China, he went to the Georgia Institute of Technology as an undergraduate and received a master's (1973) and a doctorate (1976) in electrical engineering from the University of California, Berkeley. He joined Bell Laboratories in 1976; since then, he has investigated semiconductor injection lasers and molecular-beam epitaxy of III-V compounds and has developed a number of electronic and photonic devices. In 1982 Tsang was awarded the Adolph Lomb Medal of the Optical Society of America.

Reprinted from the November 1984 issue



SEMICONDUCTORS are materials in which a small amount of energy promotes electrons from a valence band (*gray*) in which electrons are bound to atoms, to a conduction band (*color*) in which electrons move freely. Two types of semiconductor are made by introducing impurities into an intrinsic semiconductor such as silicon. In an *n*-type semiconductor (*top charts*), impurity atoms called donors each give up an electron to the conduction band. The donors themselves become positive ions. In a *p*-type semiconductor (*bottom charts*), impurity atoms called acceptors each capture an electron from the valence band, leaving a "hole." The hole, or absence of an electron, is in effect a positive charge. The donors become negative ions.



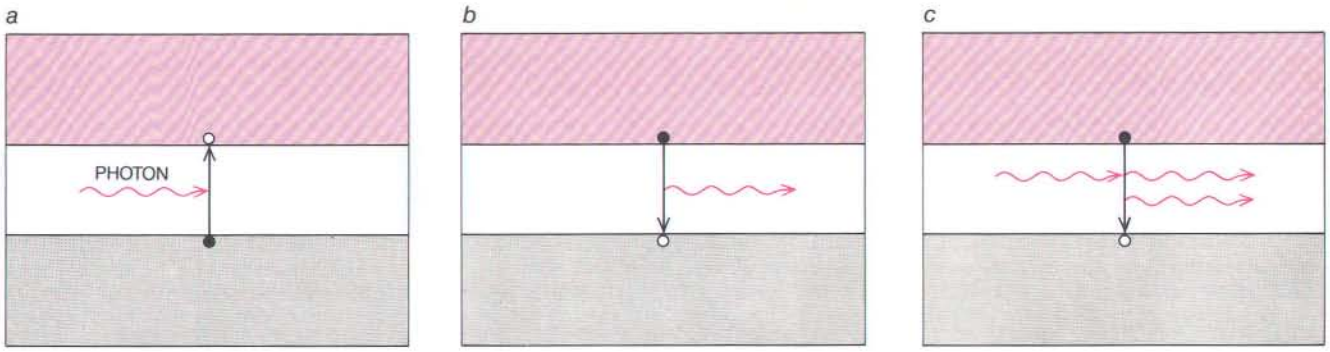
***P-N* JUNCTION is the alteration of a semiconductor material from *p*-type to *n*-type in a narrow zone. At the time the junction first arises (*top drawing*), the *n* side has an excess of electrons, the *p* side an excess of holes. Electrons and holes thus tend to diffuse. When an electron meets a hole (in an ideal semiconductor), they recombine. What remains (*bottom drawing*) are the ionized impurity atoms, which produce an electric field at and around the junction. This built-in electric field makes the *p-n* junction the central element in semiconductor electronics.**

Now suppose a pulse of light includes a range of wavelengths. Light at different wavelengths travels at different velocities in a given material. Again, therefore, the pulse of light broadens—even in a single-mode fiber. The phenomenon is called chromatic dispersion. Remarkably, the pattern of dispersion reverses itself (in present-day fibers, which are based on silica) at a wavelength of about 1.33 microns in the infrared part of the electromagnetic spectrum. Wavelengths near 1.33 microns all travel at approximately the same speed. But for wavelengths shorter than 1.33 microns the relatively "blue" part of the light travels slower than the relatively "red" part and so arrives later at the end of an optical fiber. For wavelengths longer than 1.33 microns the "blue" part arrives sooner than the "red."

A further problem limits the performance of an optical fiber. It is the optical loss in the fiber, that is, the loss of signal energy. The loss occurs when impurities in the material scatter light or absorb it. Silica glass, for example, shows a marked loss at a wavelength of about 1.25 microns and another at about 1.39 microns. Both are due to the absorption of light by hydroxyl (OH^-) ions trapped in the glass during its manufacture. Aside from these sharp losses, silica shows a broad-scale loss that rises rapidly with decreasing wavelength.

If these problems are to be minimized, the light must be carefully chosen. To avoid chromatic dispersion, a wavelength of 1.33 microns is best. To avoid optical loss, a wavelength of 1.55 microns is preferable: single-mode silica fibers are most transparent at that wavelength. One strategy would be to employ an utterly monochromatic source producing light at a wavelength of 1.55 microns; chromatic dispersion would cease to be a problem because the light consists of one wavelength, and the fiber would be at its most nearly transparent. Unfortunately, no conventional semiconductor laser currently in use emits light at a single wavelength. The pulsed output of a conventional semiconductor laser has a spread in wavelength of from five to 10 nanometers.

The lasers now serving optical communications systems are semiconductor lasers. Such devices are small (about the size of a grain of salt). They produce pulses of light from pulses of electric current; their electri-



EMISSION OF LIGHT by a semiconductor occurs in two ways: either spontaneously or under stimulation. In the spontaneous emission process electrons in the valence band have been excited into the conduction band by absorbing photons (quanta of light) whose energy exceeds the band gap: the energy difference between the bands (*a*). The excited electrons are unstable: after a short time they spontaneously "drop" to the valence band, annihilating a hole and releasing a photon whose energy equals the band gap (*b*). In stimulated emission, photons whose energy equals the band gap induce excited electrons to make their "drop" to the valence band. The photons produced by stimulated emission match the incident photons in both energy and phase, that is, the alignment of the waveform of the photons (*c*).

cal requirements are modest (a few milliamperes at one or two volts); they can generate light at the infrared wavelengths where optical fibers are most nearly transparent; and unlike, say, a gas laser, they are mechanically stable and reliable.

In order to see how the C^3 laser promises major improvements in the performance of semiconductor lasers, it is worth exploring in some detail how semiconductor lasers work. Fundamentally a semiconductor, as its name suggests, is a material whose electrical properties are intermediate between those of an insulator (in which electrons are tightly bound to atoms) and those of a conductor such as a metal (in which certain electrons—the outermost ones in the metal atoms—are free to move throughout the volume of the material). In a semiconductor the outermost electrons are bound but can be freed by means of a small amount of energy.

One way to characterize these differences is to note that in an isolated atom the amount of energy allotted to each electron (the energy "level" of the electron) is sharply delimited. In a solid, however, the atoms are arranged in a periodic lattice structure; consequently, the individual levels merge into broad bands of allowable energies separated by forbidden zones called band gaps. In a metal the outermost occupied band is only partially filled by electrons. Thus, there are vacancies available for free, or charge-carrying, electrons. In an insulator the outermost occupied band is completely filled, and a wide band gap intervenes between it and the next-higher band, which is empty. In a semiconductor the outermost occupied band is also completely filled (this band is called the valence band), but the next-higher band (called the conduction band) is only a small band gap away.

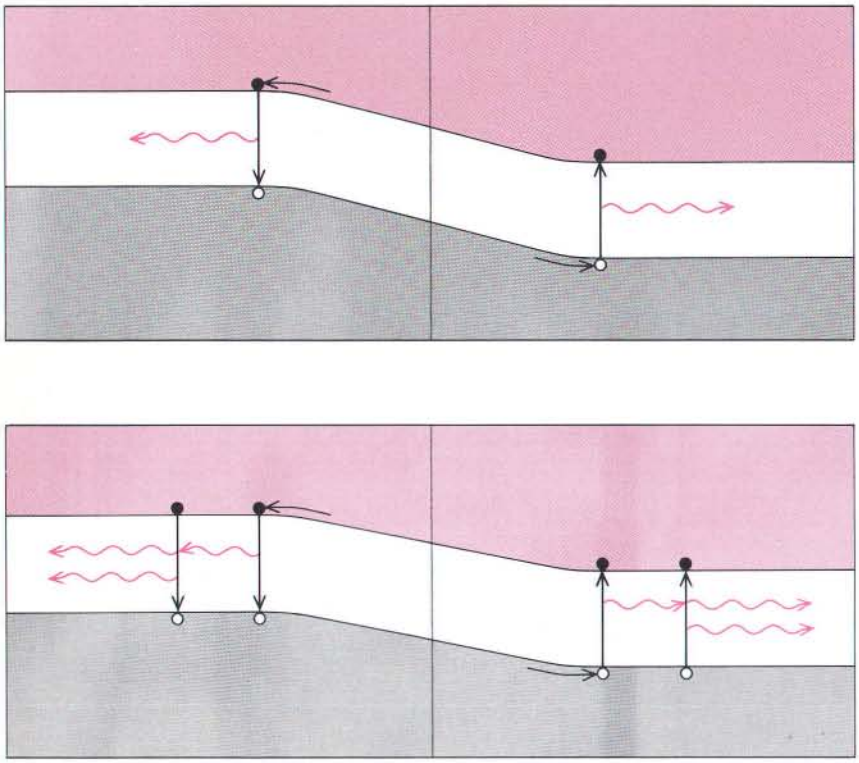
Some materials, such as the chemical elements silicon and germanium or mixtures of two different elements (say, indium and phosphorus) or three elements (say, indium, gallium and arsenic) or more, are intrinsic semiconductors. Often, however, impurities are added to intrinsic semiconductors in order to modify their electrical and optical properties. The resulting materials are termed extrinsic semiconductors. There are two kinds. In the first, termed an *n*-type semiconductor, impurity atoms called donors each contribute one of their electrons to the conduction band [*see top illustration on opposite page*]. The amount of energy required to free and "promote" this electron is quite small compared with the band-gap energy. Since the donor loses an electron, it acquires a positive charge.

In the second variety, termed a *p*-type semiconductor, impurity atoms called acceptors each capture an electron from the valence band, thereby leaving that band with a "hole": the absence of an electron. The hole acts in every respect as if it were a positive charge. Meanwhile the acceptor has acquired a negative charge.

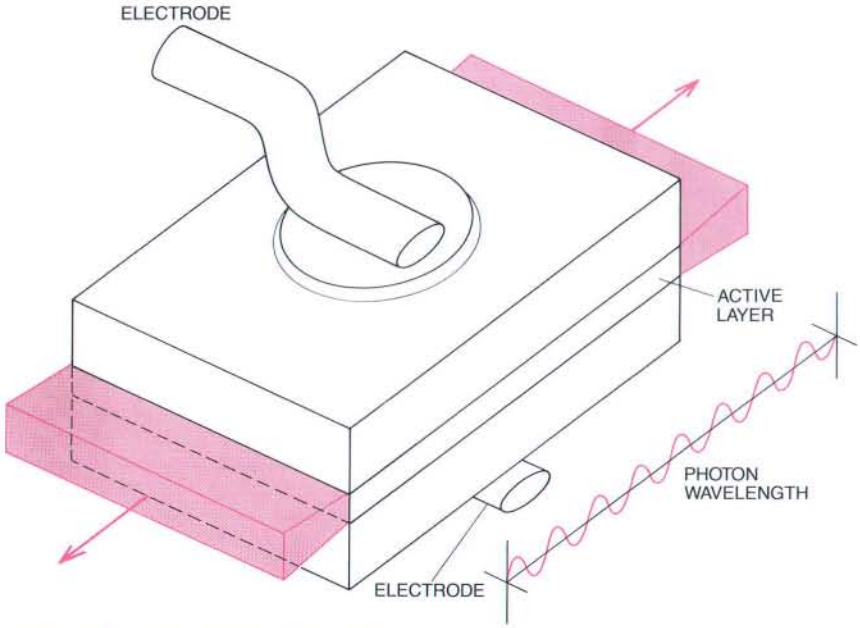
The material at the heart of modern solid state electronics is the *p-n* junction, that is, the alteration of a semiconductor from *p*-type to *n*-type material over an extremely small distance. A transistor, for example, is either a *p-n-p* material or an *n-p-n* material. In order to understand how a junction works, it is useful to imagine an instant of time in which *p* material and *n* material first come in contact [*see bottom illustration on opposite page*]. At that instant the *n* side has an excess of free electrons; the *p* side has an excess of holes. Driven, therefore, by a concentration gradient, electrons tend to diffuse toward the *p* side; holes tend to diffuse toward the *n* side.

When an electron meets a hole in an ideal semiconductor, the two recombine. As a result, charge carriers (electrons and holes) tend to disappear from the vicinity of the junction. The *p* side is left with an excess of acceptor atoms; their positions—and their negative charges—are fixed in the crystal lattice of the semiconductor. The *n* side is left with an excess of donor atoms, whose positive charges are likewise fixed in the lattice. In this way an electric field becomes established in the vicinity of the junction, a field that opposes the further diffusion of electrons or holes.

Such a device can be designed to serve as a laser. The crucial point is that three basic optical processes can occur in a semiconductor: spontaneous absorption, spontaneous emission and stimulated emission [*see illustration above*]. In the spontaneous processes a charge carrier (an electron or a hole) absorbs or emits a photon as it moves from one band to another. The energy of this photon equals that of the band gap. The wavelength of a photon is a measure of its energy; hence, the wavelength of the radiation absorbed or emitted by a semiconductor depends on the choice of a particular semiconducting material. Suppose an electron



P-N JUNCTION EMITS LIGHT under the influence of a voltage applied to the junction. Basically the voltage counters the electric field at the junction. As a result, electrons and holes diffuse across the junction and recombine, emitting photons by spontaneous emission (*top drawing*). If the applied voltage is almost large enough to nullify the electric field, stimulated emission can occur (*bottom drawing*) when electrons have been excited in sufficient number.



SIMPLEST SEMICONDUCTOR LASER is a *p-n* junction in a semiconductor crystal whose end faces are flat and perfectly parallel. The faces then form a pair of semireflecting mirrors bouncing photons back and forth through the "active" layer of the crystal. When current is injected, photons arise by spontaneous emission. The ones traversing the semiconductor bring on an avalanche of stimulated emission. Reflections at the mirrors are self-reinforcing if the wavelength of the photon fits evenly into the length of the laser. An example is at the right.

in the valence band absorbs a photon whose energy is greater than the band gap. The electron will be excited into the conduction band, leaving a hole in the valence band. Conversely, suppose an electron is in the conduction band. It is in an excited state; in a word, it is unstable. After a short time, and without any external stimulus, it will make a transition into the valence band, where it will annihilate a hole and release a photon whose energy precisely equals that of the band gap.

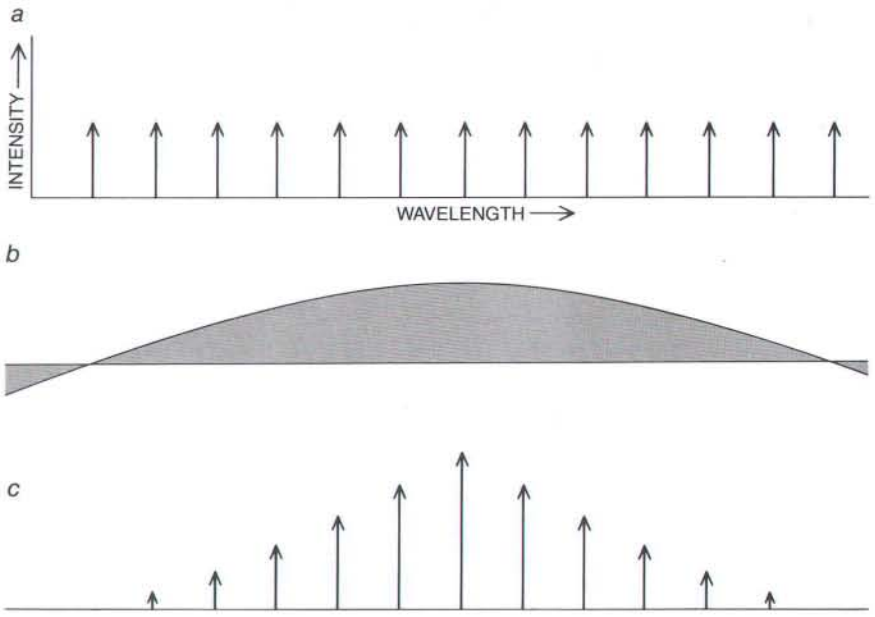
The third basic process—stimulated emission—is essential to laser action. It takes place when a photon whose energy precisely equals that of the band gap impinges on electrons in the conduction band. In this situation an electron in the conduction band will be induced to make its transition to the valence band and so emit a photon. Remarkably, the incident photon and the photon emitted by the electron will match each other not only in energy but also in phase. That is, the crests and troughs of their waves will align. It should be said that photon emission can arise from transitions between a band and the energy level occupied by an impurity atom and even from transitions between two impurity levels. Moreover, the energies of charge carriers in the conduction band and valence band obey a statistical distribution. Further still, the energy levels of impurity atoms form narrow energy bands. For all these reasons, the electromagnetic radiation emitted by a semiconductor includes a narrow range of wavelengths rather than the single wavelength defined by the band gap.

A *p-n* junction can be an excellent light emitter; one need only connect the junction to a supply of electric current. The current places a voltage across the junction—a voltage that partially counters the electric field intrinsic to the junction. In this way, the energy barrier raised against the flow of electrons and holes is reduced [*see top illustration at left*]; hence, electrons move from the *n* side to the *p* side, where they recombine with holes, emitting photons by spontaneous emission. A similar process produces photons when holes move from the *p* side to the *n* side and recombine with electrons there. If the applied voltage is almost large enough to flatten the energy barrier, great numbers of charge carriers become capable of surmounting the barrier; the resulting flow of carriers leads to a carrier distribution called

a population inversion. More important, the number of photons emitted in the vicinity of the junction, in what is called the active layer of the semiconductor crystal, becomes so great that stimulated emission occurs. When the current is turned off, the emission of photons stops. Hence, the semiconductor emits pulses of light in response to pulses of current: it is a transducer, or converter, of electric signals into optical signals.

One further condition must be fulfilled for the device to be a laser. It is that two end faces of the crystal perpendicular to the *p-n* junction must be flat and perfectly parallel to each other [*see bottom illustration on opposite page*]. The faces then form a pair of semireflecting mirrors that bounce photons generated at the junction back into the active layer. As the photons traverse the region, they induce an avalanche of stimulated emission. In short, they amplify the light. (The word "laser" is an acronym for "light amplification by stimulated emission of radiation.") Photons escaping the semiconductor through the semireflecting mirrors form the laser beam.

The semireflecting mirrors are really a modern version of an optical apparatus invented in 1896 by the French physicists Charles Fabry and Alfred Perot. Hence, the limitations that apply to the apparatus (now called a Fabry-Perot resonant cavity) also apply to the laser. In particular, the light reflected by the mirrors can reinforce itself only if the light and its reflection are in phase. To put it another way, the wavelength of the light must fit evenly into the length of the laser—or rather the effective length: the measured length multiplied by the refractive index of the material between the mirrors. Wavelengths that satisfy the condition are called resonant wavelengths. All other wavelengths are suppressed. In principle, the resonant wavelengths are infinite in number. But since a particular *p-n* junction generates light in only a narrow range of wavelengths (usually called the gain profile of the laser), the beam emitted by the laser will consist of resonant wavelengths that fall within the range. The end faces of a semiconductor laser are typically 200 to 400 microns apart. If the laser is designed to emit light at a wavelength of 1.55 microns, it will actually emit a number of resonant wavelengths that differ by "mode spacings" of approximately one to two nanometers. Once the material of the active layer has been chosen, the wavelengths are unchangeable.



ENERGY DIAGRAM for the simplest semiconductor laser shows why its output beam jumps randomly among several wavelengths. In principle, the laser resonates at the infinite number of wavelengths that fit evenly into the length of the laser (*a*). The *p-n* junction, however, produces photons only in a narrow range of wavelengths called the gain profile (*b*). Thus, the beam emitted by the laser includes only the resonant wavelengths positioned within the profile (*c*).

Physics alone does not govern the nature of semiconductor lasers. Some engineering practicalities also have a bearing on the design. For one thing, the useful life of a semiconductor laser depends on limiting the amount of current applied to the device. Several methods are employed. In one method, further chemical elements are introduced into the crystal so that the active layer of the laser is between regions where the band gap is particularly wide. The resulting "heterostructure" has built-in energy barriers that help to prevent the diffusion of charge carriers out of the active layer. Stimulated emission can then be attained at relatively low values of applied current. The active layer can in fact be made a narrow tube surrounded on all sides by wide band-gap materials. The resulting devices are called buried-heterostructure lasers.

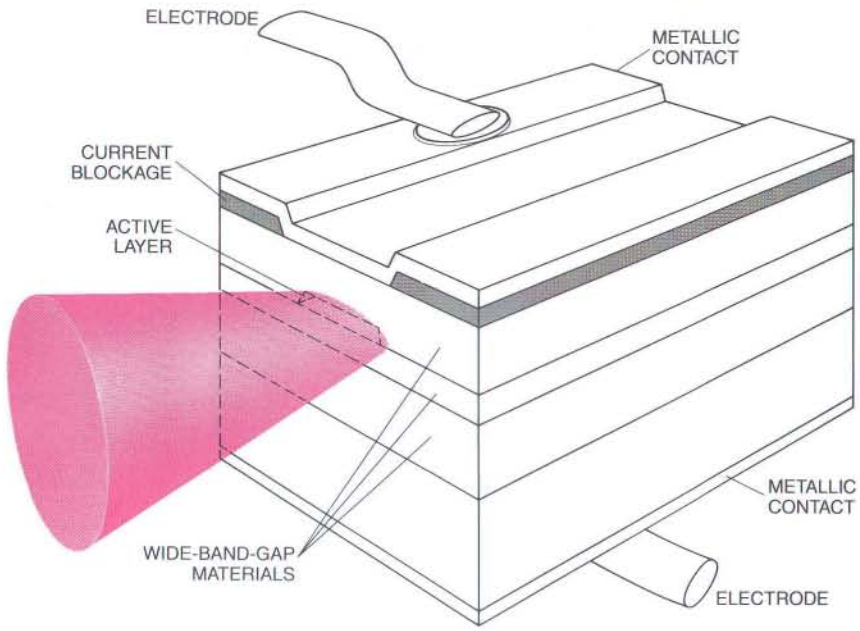
Nature has been kind: the refractive index in the active layer turns out to be greater than that of the surrounding wide band-gap material. As a result, the tubelike active layer in a buried-heterostructure laser acts like the core of an optical fiber: light generated in the active layer is guided along the tube. The guiding is valuable because the process of stimulated emission requires the interaction of photons and excited charge carriers. Together the advances in the design of semiconductor lasers and the development of excellent crystal growth and fabrication techniques now yield lasers that can operate at room temperature for more than a million hours on currents as low as 2.5 milliamperes at one or two volts.

This is not to say the buried-heterostructure laser is ideal for optical communications. For one thing, the beam of even a buried-heterostructure laser jumps randomly among the available resonant wavelengths. The jumping, called mode partition, is particularly troublesome when the laser is turned on and off rapidly in an effort to code information as pulses of light. The jumping cannot be neglected in a modern optical communications system, where the tolerable error rate is one bit in 10^9, or even one in 10^{11}. Suppose two pulses of light emitted by a laser designed to operate at about 1.55 microns (1,550 nanometers) actually differ in wavelength by two nanometers (the mode spacing for a laser cavity length of 250 microns). The pulses travel 100 kilometers in a single-mode optical fiber whose core is made of silica. Because of chromatic dispersion,

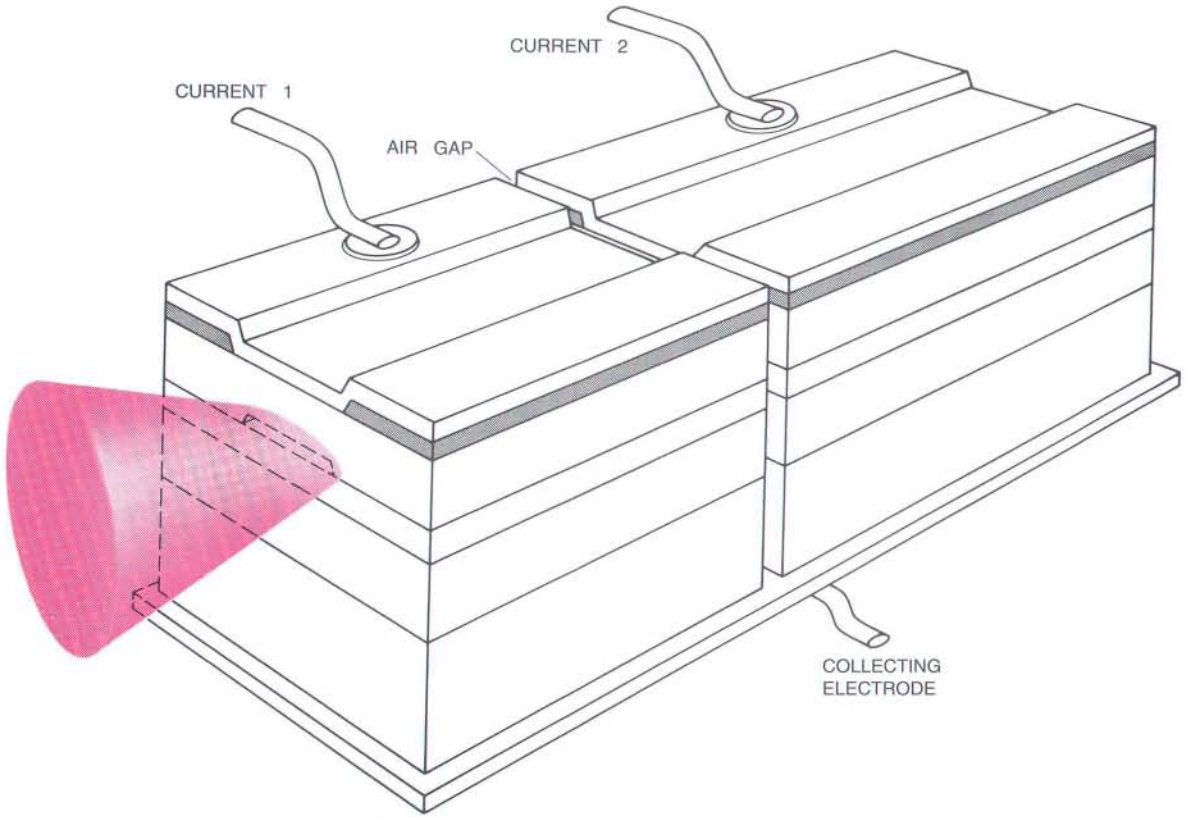
the transmission time for the relatively blue pulse (the one at 1,548 nanometers) will be 3.5 nanoseconds less than the transmission time for the relatively red pulse (the one at 1,552 nanometers). As a result, the information-carrying capacity of the system will be limited to a maximum of 150 million bits per second; otherwise the pulses would smear over one another. Hundreds of millions of bits per second may seem to be a large number. Actually the goal for modern optical communications lines lies well within the billion-bit-per-second range.

It was the pursuit of ultrahigh-capacity, long-distance optical communications systems that led to the C^3 laser, a device that not only emits the purest known laser beam when the device is being pulsed at rates as high as two billion pulses per second but also is the first single-wavelength laser that can be tuned electronically over a wide range of output wavelengths.

The C^3 laser consists of two semiconductor lasers that differ slightly in length, say, by 20 percent. Because they



BURIED-HETEROSTRUCTURE LASER resulted from efforts to improve the performance of the basic semiconductor laser. It reduces the *p-n* junction to a "tube" running the length of the semiconductor crystal. Then it surrounds the tube with layers of semiconductor whose wide band gap raises an electrical barrier confining charge carriers within the tube. The wide band-gap material also confines the photons produced at the junction. The laser beam spreads because of diffraction occurring where the beam emerges from the face of the device.

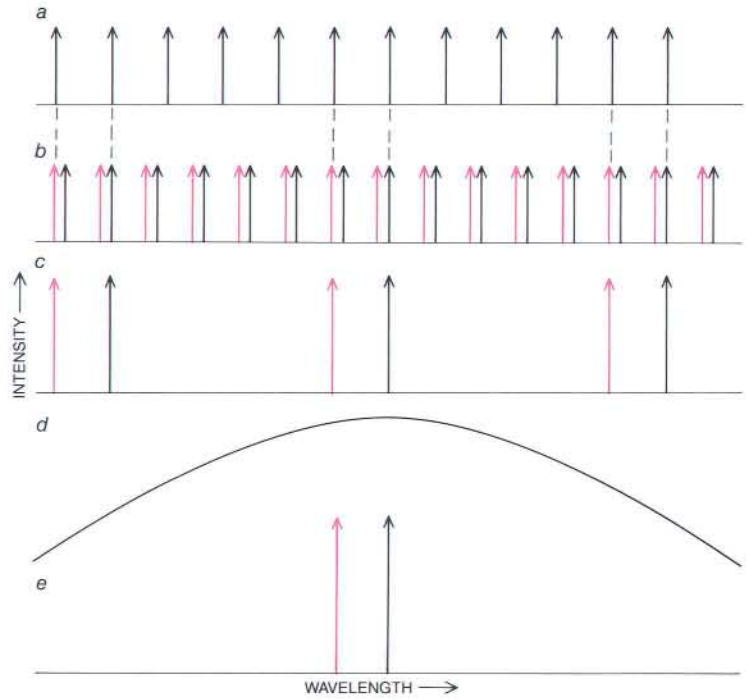


ALIGNMENT OF TWO LASERS composes the C^3 laser. The half lasers have different lengths; hence, their resonant wavelengths are differently spaced, and only a few of them match. The mismatches are suppressed. Among the matches, moreover, only one is near the peak gain. Thus, the beam of a C^3 laser consists of that wavelength almost exclusively. The probability of the beam's jumping to another wavelength is less than one in 10 billion beam samplings.

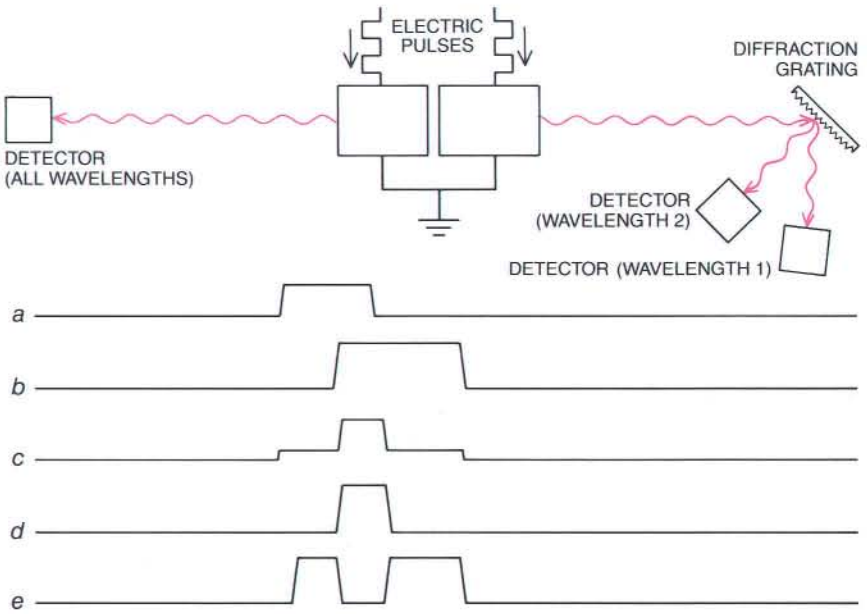
differ in length, the spacing of their resonant wavelengths differs. The two lasers are positioned close to each other, with their active layers aligned, so that the beam emitted by each laser enters the other [*see bottom illustration on opposite page*]. The light emitted by one is then suppressed by the other, unless the wavelength is resonant in both. In general the wavelengths will not match; indeed, the output beam will be confined almost perfectly (nothing in quantum mechanics is absolutely certain) to a single wavelength: a matching resonance near the peak of the lasers' gain profile [*see top illustration at right*]. In this way, mode partition is eliminated.

How is the C^3 laser tuned? By injecting current into one of the individual lasers. A central point about the C^3 laser is that the two lasers in it are optically coupled but electrically isolated: each receives its own electric current. Suppose, then, one of the lasers is given a current that places it above its lasing threshold (the minimum current that produces a population inversion and stimulated emission). The positions of its resonant modes will then be fixed. Meanwhile the second laser is given a current that keeps it below its threshold. Under this condition the second current serves exclusively to control the density of charge carriers in the active layer of the second laser. (If the second laser were lasing, the application of electric current would serve only to increase the stimulated emission of photons.) The density of charge carriers turns out to govern the refractive index of the laser. In turn the refractive index governs the effective length of the laser, and the effective length governs the resonant wavelengths. In sum, the varying current applied to the second laser serves to shift the resonant wavelengths of the second laser with respect to those of the first. As a result, the resonances that matched will fall out of alignment, and other matches will take their place.

At Bell Laboratories we have tuned a 1.55-micron C^3 laser to 15 different infrared wavelengths in a range of 30 nanometers. Each successive tuning required a change in current to the subthreshold laser amounting to little more than a milliampere; the amount of time the switching required was about a billionth of a second. Mode partition was virtually absent: it occurred less than once in 10 billion samplings of the beam. (Modern ultrahigh-capacity, very long distance optical communica-



TUNABILITY OF A C^3 LASER by current injection is shown in the laser's energy diagram. The current to one of the half lasers puts it above its lasing threshhold, so that its resonant modes are fixed (*a*). The second half laser is kept below threshold. Still, it has resonant modes (*black lines in* b), and some of them (*black lines in* c) match the modes of the light-emitting half laser. The match at the peak of the laser's gain profile (*d*) determines the wavelength of the C^3 laser's output (*e*). Now the current applied to the second half laser is changed. The change alters its resonant modes (*colored lines in* b). Hence, a new set of matching resonances is established (*colored lines in* c) and with it a new output wavelength (*colored line in* e).



OPTICAL LOGIC CIRCUIT may be a future application for the C^3 laser. The circuit arises from the laser's tunability and from the electrical isolation of the two half lasers composing a C^3 laser. Independent trains of electric pulses (*a, b*) are applied to the half lasers. Simultaneous pulses cause the emission of light at wavelength *1*. A pulse to one of the half lasers causes the emission of light at wavelength *2*. The detection of light at both wavelengths amounts to the logic operation designated *or* (*c*).The detection of wavelength *1* amounts to the logic operation *and* (*d*). The detection of wavelength *2* amounts to the logic operation *exclusive or* (*e*).

tions systems require an error rate of less than one incorrect bit in a billion bits of transmitted information.) The laser was maintained at any of its wavelengths by holding constant the current applied to the second laser. To generate optical pulses, the first laser was maintained above lasing threshold and given trains of electric pulses. At a certain strength the application of electric current to the second laser placed it above its lasing threshold. As a result, the tunability ceased. The C^3 laser then functions at a single wavelength in spite of further increases in the applied electric current.

A C^3 laser is made by cleaving a conventional semiconductor laser along a crystal plane parallel to its end faces. The result is two shorter lasers. In isolation each of the latter would be a conventional laser. In fact, the two lasers are not isolated. One of the surfaces of the parent laser, a surface parallel to the *p-n* junction, is coated with a gold pad about three microns thick. The pad resists the cleaving; thus it serves as a hinge holding the half lasers together with their active layers in precise alignment and an air gap a few microns wide between them. Contrary to the reservations expressed when the technique was tried, the width of the air gap and the precise difference in length between the two half lasers turn out not to affect the performance of the device in any critical way. The technique is applicable to the manufacture of lasers made from a wide range of materials emitting at wavelengths from the visible (at about 0.7 micron) to the far infrared (at about 30 microns).

Among the potential uses for the C^3 laser, three forms of optical communications strike me as having particular promise. First, the extraordinarily monochromatic output of a C^3 laser eliminates the problem of chromatic dispersion in optical fibers and so facilitates the transmission of digital information in a single-mode fiber at a wavelength of 1.55 microns: the wavelength at which a silica fiber is most nearly transparent to electromagnetic radiation. In one test of the laser's ability at 1.55 microns, Bell Laboratories has transmitted digital information in an optical fiber more than 120 kilometers long at a rate of one gigabit (10^9 bits) per second without reamplification along the way. The frequency of error was less than two bits in 10^{10}. At such a rate, one could transmit the text of the *Encyclopaedia Britannica* in less than half a second, and the text would be received virtually without error (one letter or punctuation mark might be a misprint). In a further test, also done at Bell Laboratories, digital information was transmitted 160 kilometers at a rate of 420 megabits (420×10^6 bits) per second without reamplification. The error rate was less than five bits in 10^{10}.

The second application for C^3 lasers in optical communications lies in wavelength-division multiplexing. Here the output beams of several C^3 lasers, each laser tuned to a different wavelength, are coupled to a single optical fiber. The fiber can then carry several independent messages.

The third application enhances still further the information-carrying capacity of a single optical fiber. At rates on the order of a billion switchings per second, one can shunt the output wavelength of a C^3 laser among as many as 15 modes spaced about two nanometers apart. Thus the single-wavelength transmission of information, with high-power and low-power pulses representing the binary digits 1 and 0, respectively, yields to multiwavelength transmission. In the simplest possibility two different wavelengths are employed to represent 1 and 0. A more complex scheme calls for switching between four different wavelengths. A single pulse would then signify the binary data 00, 01, 11 or 10, depending on the wavelength. In this way, each pulse would carry two bits of information. Switching among eight wavelengths would enable a single pulse to carry three bits of information.

Beyond the possibilities in optical communications are further possibilities in which C^3 lasers perform the switching and logic operations now done by electronic means. The scheme relies on the electrical isolation of the two half lasers composing a C^3 laser. Each half laser receives a train of electric pulses [*see bottom illustration on preceding page*]. If a pulse of current arrives simultaneously at each half laser, the C^3 laser emits a pulse of light at a certain wavelength; call it wavelength 1. If a pulse of current arrives at one half laser but not at the other, the C^3 laser emits a pulse of light at a different wavelength, wavelength 2. If no pulses of current arrive, no pulses of light are produced.

The pulsed beam of light emerging from one end of the C^3 laser is directed into a photodetector sensitive to all wavelengths. Accordingly, the photodetector signals whenever one half laser or both half lasers receive a pulse of electric current. The detector is performing the logic operation known as *or*. Meanwhile the pulsed beam emerging from the other end of the C^3 laser is directed into a diffraction grating, which splits light into a fan of wavelengths. A pair of photodetectors can then detect wavelength 1 and wavelength 2 independently. The presence of wavelength 1 signals that both half lasers are receiving a pulse of electric current; thus, the wavelength-1 detector performs the logic operation *and*. The presence of wavelength 2 signals that one half laser or the other—but not both—is receiving a pulse of electric current. This is the logic operation *exclusive or*.

The scheme has several advantages. It produces multiple logic outputs for a single pair of electric inputs. It is fast: the switching time for a C^3 laser is as short as a nanosecond, so that information processing in the gigabit-per-second range is conceivable. Finally, the scheme yields optical output from electric input; hence, it can serve as an electronic-optical transducer. Applications for optical logic and switching may lie some time ahead. Still, it is clear that semiconductor optics can serve in that capacity. The C^3 laser is a straightforward extension of the conventional semiconductor laser, yet its characteristics and capabilities may prove to be important in many different ways.

FURTHER READING

SINGLE FREQUENCY INJECTION LASER DIODES FOR INTEGRATED OPTICS AND FIBER OPTICS APPLICATIONS. L. B. Allen, H. G. Koenig and R. R. Rice in *Proceedings of the Society of Photo-Optical Instrumentation Engineers,* Vol. 157, pages 110–117; 1978.

MONOLITHIC TWO-SECTION GaInAsP/InP ACTIVE-OPTICAL-RESONATOR DEVICES FORMED BY REACTIVE ION ETCHING. L. A. Coldren, B. I. Miller, K. Iga and J. A. Rentschler in *Applied Physics Letters,* Vol. 38, No. 5, pages 315–317; March 1, 1981.

HIGH-SPEED DIRECT SINGLE-FREQUENCY MODULATION WITH LARGE TUNING RATE AND FREQUENCY EXCURSION IN CLEAVED-COUPLED-CAVITY SEMICONDUCTOR LASERS. W. T. Tsang, N. A. Olsson and R. A. Logan in *Applied Physics Letters,* Vol. 42, No. 8, pages 650–652; April 15, 1983.

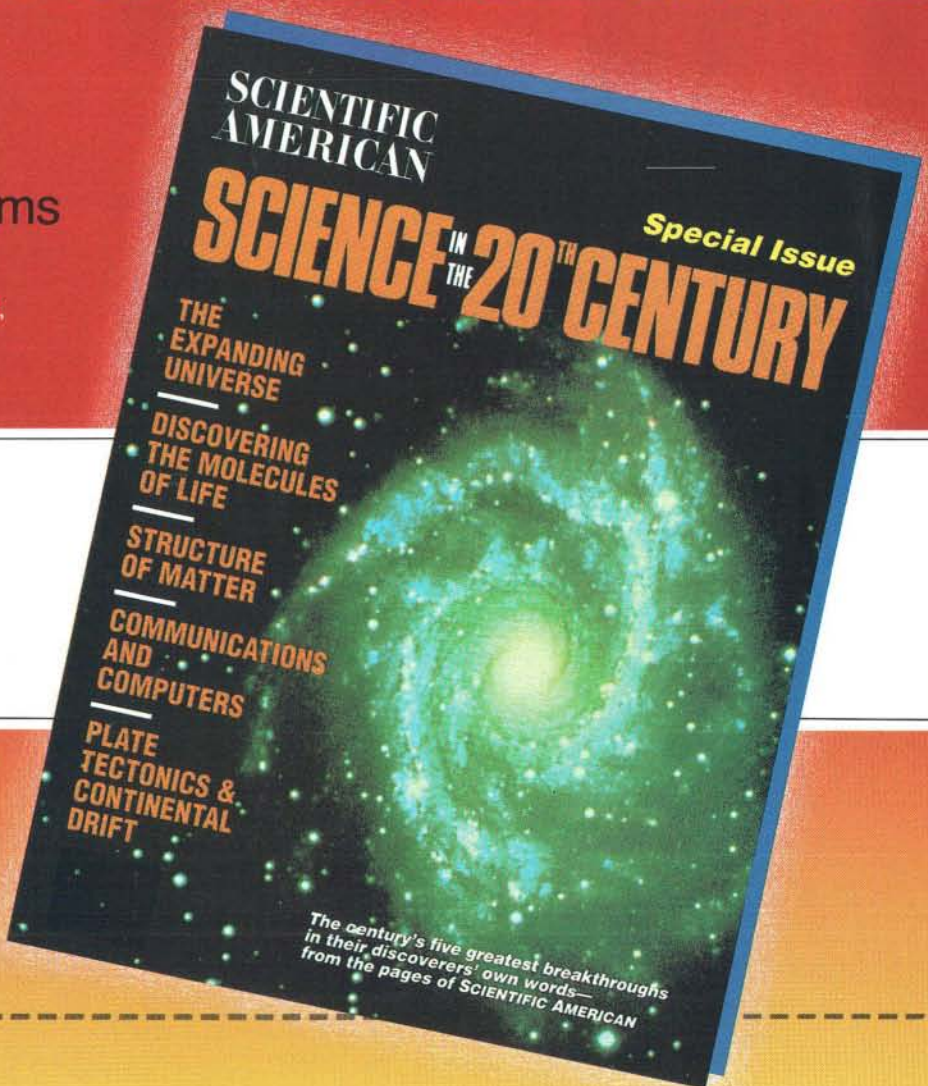
SCIENTIFIC AMERICAN

PRESENTS A SPECIAL ISSUE

Sponsored by

AT&T Network Systems

THE FIVE MAJOR ADVANCES OF SCIENCE, AS REPORTED IN THE PAGES OF *SCIENTIFIC AMERICAN*, IN THE ACTUAL WORDS OF THE SCIENTISTS WHO MADE THEM HAPPEN.



NOW AVAILABLE IN BULK FOR INDUSTRY AND EDUCATION

The discoveries reported by these authors are both pivotal and fascinating. The inflationary universe, the structure of genes, the essence of matter, transistors, lasers, the earth's hot spots, and more—all major scientific breakthroughs in the 20th Century and each one revolutionizing society, the economy, the way we know ourselves. Authoritatively written by the scientists themselves.

☑ YES, I would like to receive the special issue SCIENTIFIC AMERICAN *Science in the 20th Century*. $3.95 each. Only $3.00 for orders of 50 or more copies.

Please send _____ copies @ $3.00 $_____
(Minimum order 50 copies)

I'm not ordering 50 copies now. Please send me _____ copies @ $3.95 $_____
(Add $1.00 per copy for postage and handling on orders of less than 10 copies)

TOTAL ENCLOSED $_____
Foreign orders, please add $1.00 per copy.

Please make check or money order payable to: SCIENTIFIC AMERICAN.

Send order to: SCIENTIFIC AMERICAN
Dept. 20C
415 Madison Avenue
New York, NY 10017, U.S.A.

Name _____

Company _____

Address _____ Apt. _____

City _____ State _____ Zip _____

STC B

NEW BULK RATES. SAVE 24%

The Connection Machine

Most computers have a single processing unit. In this new parallel computer, 65,536 processors work on a problem at once. The resulting speed may transform several fields, including artificial intelligence

by W. Daniel Hillis

In the past three decades remarkable changes have taken place in digital computers. The amount of computational power that once required a room full of vacuum tubes can now be found in hand-held devices. Complex computations that would once have taken days to perform can now be done in seconds. Yet in certain fundamental respects the design of the digital computer remained unchanged between the days of the ENIAC (one of the first large-scale digital machines, built at the University of Pennsylvania in the late 1940s) and the current generation of supercomputers. Most modern computers—from supercomputers to microprocessors—are similar to the ENIAC in that the memory and the central processing unit are separate entities. For a computation to be performed, the appropriate data must be retrieved from the memory and brought to the central processor; there it is operated on before being returned to the memory.

Such a design is called sequential because the processing operations are performed one at a time. The sequential design was adopted mainly for utilitarian reasons. In the early days of digital computing the memory and the central processing unit were made of different materials. Since memory was cheaper than processing, it was desirable to maximize the efficiency of the processing unit at the expense of the memory's efficiency. And that is just what the sequential design does. Today, however, the memory and the central processor are fabricated from the same etched silicon wafers. In a typical computer, more than 90 percent of the silicon is devoted to memory. While the central processor is kept wonderfully busy, this vast majority largely sits idle. At about $1 million per square meter, processed packaged silicon is an expensive resource to waste.

Clearly, the general solution to this problem is to find a way to unify processing capacity and memory. But how? One answer is to exploit many small processors, working simultaneously, each accompanied by a small memory of its own. In such a design, which is called parallel processing, memory capacity and processing capacity can both be utilized with high efficiency. This is the approach my colleagues and I have taken in building a parallel computer called the Connection Machine. The machine contains 65,536 simple processors. Each processor is much less powerful than a typical personal computer, but working in tandem they can execute several billion instructions per second, a rate that makes the Connection Machine one of the fastest computers ever constructed.

Yet the most interesting thing about the Connection Machine is not its brute speed but its flexibility. Special-purpose devices have been built that exploit parallelism to perform specific tasks quite quickly. Like idiots savants, however, such machines are usually quite awkward outside their specialties. In contrast, the Connection Machine can operate at its peak processing rate in a wide range of applications. The key to such flexibility is a communications network that enables the multitude of processors to exchange information in the pattern best suited to the problem at hand. The Connection Machine is not just a prototype. About a dozen Connection Machines are already in commercial use, and they have begun to change the way digital computing treats problems in physics, image processing, text retrieval and even artificial intelligence.

In order to understand the benefits of parallel processing, it is helpful to think about the difference between the way a conventional computer deals with an image and the way the same image is treated in the human brain. From the pair of two-dimensional images falling on the retinas, a human being is able—without apparent effort—to reconstruct a three-dimensional model of the world and maintain that model as the two-dimensional images change rapidly. Computers can be programmed to carry out part of the task, but even quite fast computers take hours to do what the human brain can do in fractions of a second [see "Vision by Man and Machine," by Tomaso Poggio; SCIENTIFIC AMERICAN, April 1984]. The brain maintains its advantage in spite of the fact that its components—neurons—are apparently millions of times slower than the computer's transistors.

Why, then, is the brain so much faster than the computer? The visual circuitry of the brain is not fully understood, but it is clear that in some areas of the brain the principles of parallel processing are at work. In those parts of the brain the entire image is processed at once. The computer, however, examines the image one tiny spot at a time, as if it were looking through a minute keyhole. In the computer the image is represented as an array of

W. DANIEL HILLIS is founding scientist at the Thinking Machines Corporation in Cambridge, Mass., and the architect of the Connection Machine system. He graduated from the Massachusetts Institute of Technology in 1978 and obtained his master's degree there in 1982. In 1985 he won an ACM Distinguished Dissertation Award for his doctoral thesis, which he undertook at the M.I.T. Artificial Intelligence Laboratory. Hillis is the author of *The Connection Machine* and many articles on robotics, artificial intelligence and systems architecture.

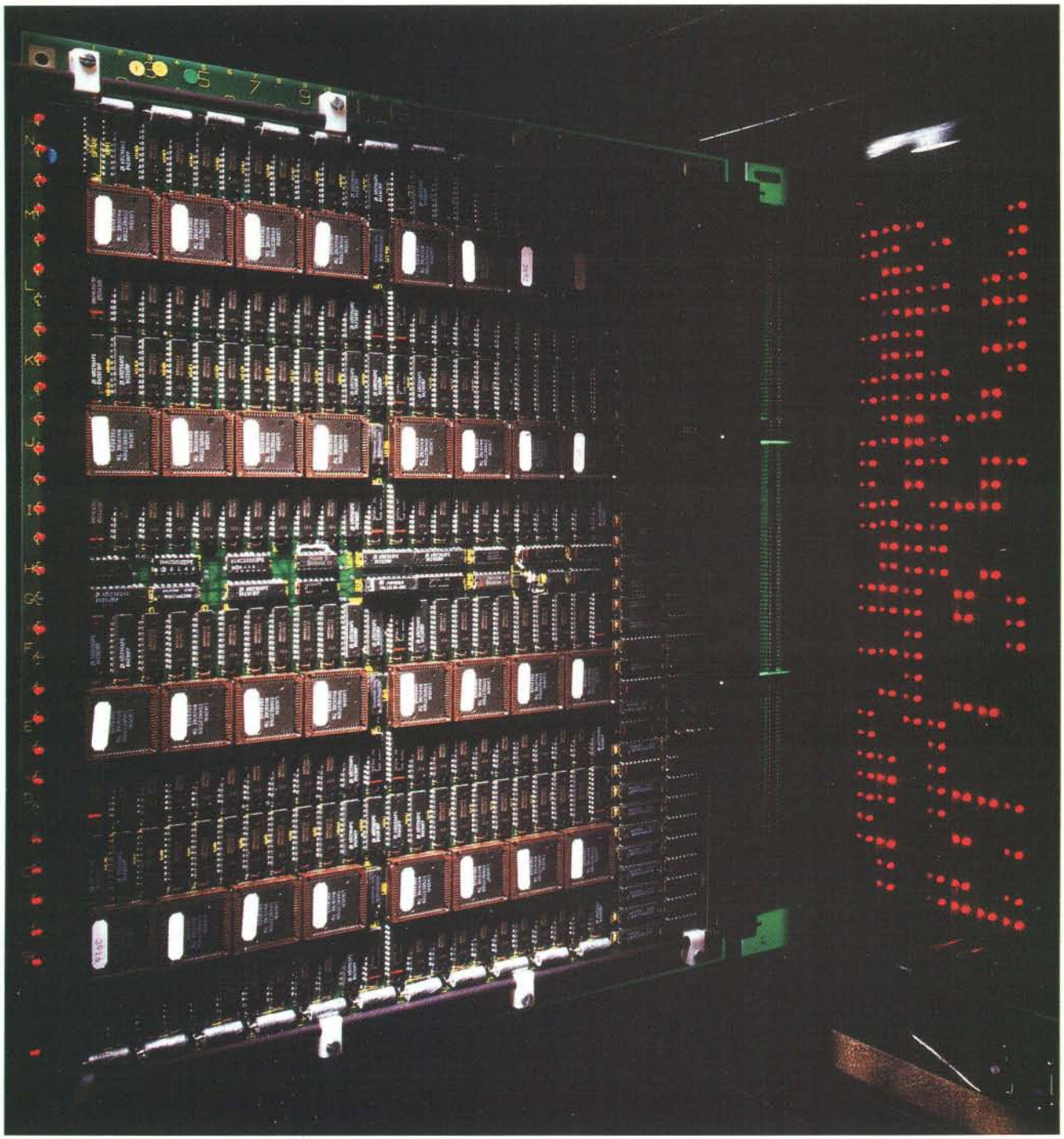
Reprinted from the June 1987 issue

CONNECTION MACHINE is a cube 1.5 meters on a side made up of eight subcubes, each containing 16 boards arranged vertically. On each board are 32 custom chips. Every chip includes 16 processors, each with a small amount of memory. The lights indicate the status of the chips; they are for troubleshooting. Operating in parallel, the 65,536 processors can execute several billion instructions per second, making the Connection Machine one of the fastest computers ever built.

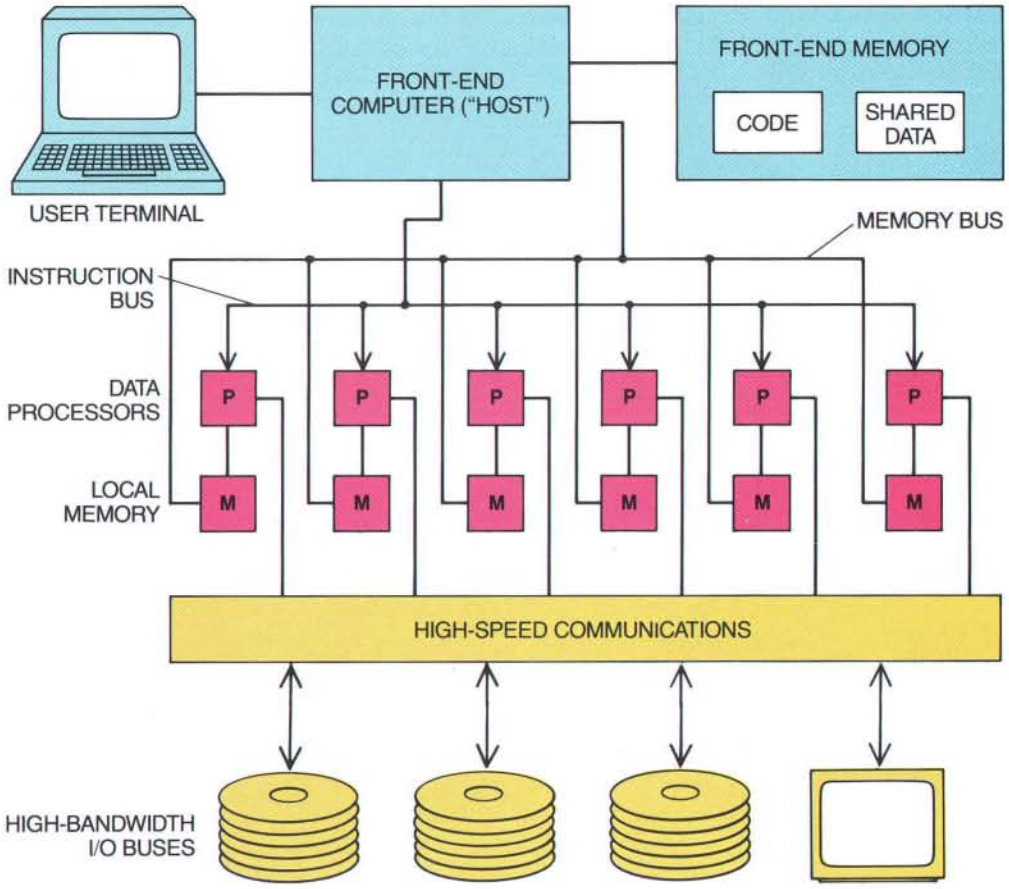
numbers, each of which corresponds to the intensity of the light at a particular point. A typical low-resolution array might be a square with 256 points on a side. A conventional computer operates on only one of the square's 65,536 points at a time. Hence, even a simple image-processing operation includes 65,536 steps.

The Connection Machine, on the other hand, assigns a single processor to each point of the image. Since every operation can be performed on all the points simultaneously, a calculation involving the entire image is as fast as a calculation involving only a single point. For example, to find all the points in the image that are brighter than a certain minimum, a sequential machine must check the 65,536 tiny elements in succession, comparing each one with the threshold value. In the Connection Machine that comparison is made simultaneously by the 65,536 processors—each one operating on a single element of the image.

The threshold comparison is particu-



BOARD slides out of the Connection Machine. The square objects are the chips, each with its 16 processors. The rectangular objects include memory units and devices for routing communications among the assembled processors.



SYSTEM DIAGRAM shows the manner in which the Connection Machine operates in association with a conventional computer, which is called the host. A user of the system interacts with the host by means of a conventional computer language that has been modified for parallel programming. Rather than carrying out repetitive operations one at a time, however, the host computer delegates them to the Connection Machine, where the operations are done in parallel. The results of these operations can be obtained by various input-output devices, including high-resolution visual displays.

larly simple because it can be carried out independently by each processor. Most interesting computations, however, require that the processors exchange information as the operation proceeds. Consider the common image-processing operation called convolution. Convolution blurs an image by averaging each point with its nearest neighbors in the two-dimensional grid. (Convolution, which is analogous to operations carried out in the human visual system, is useful for removing insignificant details and bringing out significant objects.)

To complete a convolution, each processor must read a value from the processors that store information about the points to the left, right, above and below the point in question. In effect the processors must "talk" to one another. One way to accomplish such a pattern of communication is to wire the processors in a two-dimensional grid. Since each processor would be wired to its four nearest neighbors, the grid corresponds directly to the communication paths required for convolution. Indeed, some parallel computers specialized for image processing are wired in a two-dimensional grid. That pattern works well for convolution but not for other computations.

For example, in computing the average intensity of all points in the image, a pattern of connections resembling an inverted tree is the most convenient. The average intensity of an image containing 65,536 points can be calculated by first computing the average of every pair of points, then the average of each pair of pairs, and so on. In 16 steps the average can be derived. In its last few steps the computation requires an exchange of information about points that are widely separated in the image; therefore, the two-dimensional grid is not a convenient pattern of wiring.

The general principles to be derived from these examples are that each type of computation may require its own pattern of connections and that each processor may need to communicate with any other. Therefore, in designing the Connection Machine, we chose a communications network in which any processor can communicate with any other. As a result of such flexibility, the programmer is free to choose the algorithm that is most appropriate for solving the problem at hand without having to worry about the limits imposed by the pattern of wiring.

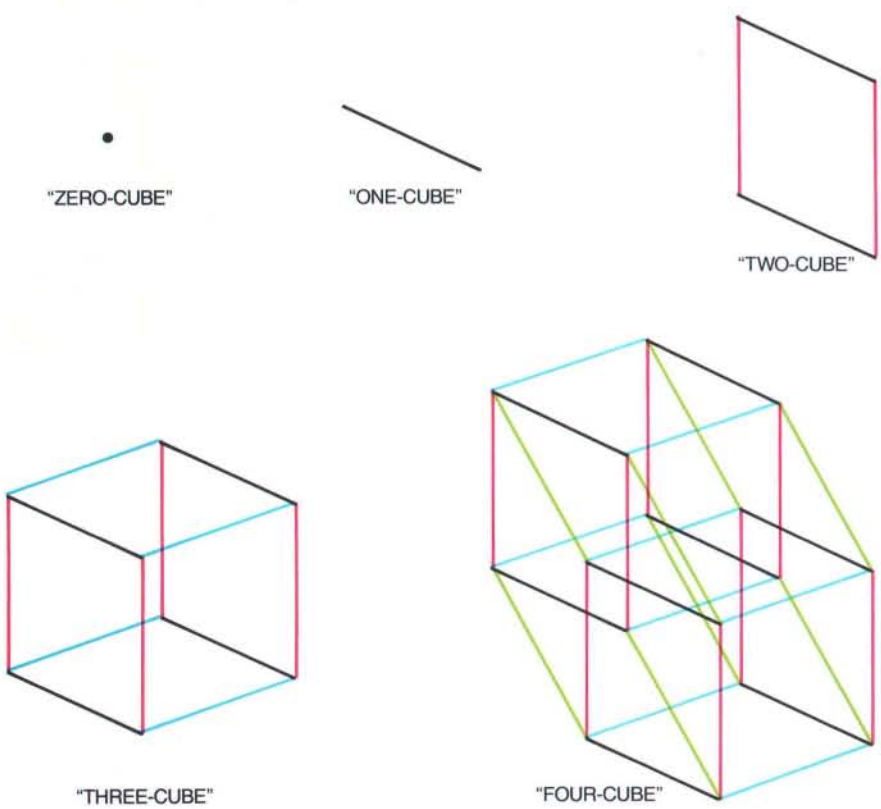
The basic replicated unit of the Connection Machine is an integrated circuit consisting of 16 small processors and a device for routing communications. Each of the processing units is associated with 4,096 bits of memory. (A typical personal computer has 256,000 bits or more.) The 16 processors are etched on a single chip, and 32 of these chips are packaged on a single printed-circuit

board. There are 128 such boards in the machine, and they are arranged in a cube 1.5 meters on a side. For purposes of troubleshooting, each chip is connected to a light on the edge of its board; the array of lights forms a pattern on the face of the cube as the machine operates.

The 16 processors on each chip are connected by a switch that makes it possible to create a direct connection between any pair of processing units. Implementing such direct connections between every pair of processors among the 65,536 in the system would require more than two billion wires, obviously an impractical figure. Instead the routing device on each chip is connected to 12 other routers in the system. The routers are wired according to a pattern called a Boolean *n*-cube. The *n*-cube is a generalized version of an ordinary three-dimensional cube that has some excellent properties as a network for communications among processors.

The full mathematical detail of the *n*-cube is somewhat beyond the scope of this article, but its general principles are not difficult to grasp. One can imagine an ordinary three-dimensional cube as one member of a series of "cubes" corresponding to different spatial dimensions. For example, a line segment might be thought of as a "one-cube," or a cube in one dimension. Joining two one-cubes by their ends yields a two-cube, or a square. Joining two two-cubes by their corners yields a three-cube, which is what we ordinarily think of as a cube. Similarly, joining two three-cubes by their corners yields a four-cube [*see illustration below*]. The process may be repeated any number of times, and it can readily be shown that a 12-cube has 2^{12} (4,096) corners, or one for each chip in the Connection Machine.

Such a Boolean *n*-cube is a valuable arrangement for several reasons. In the first place, no processor in the 12-cube is more than 12 wires away from any other, which facilitates communication in the network. Second, the design of the *n*-cube accords well with the binary logic of the computer. In the digital computer, all data are stored as strings of bits, each with a value of either 0 or 1. Now, each cube in the *n*-cube has two subcubes, which may be designated 0 and 1, respectively. As a result, each point in the *n*-cube has a unique address specified by a string of 12 binary bits. The first bit specifies which of the 11-cubes within the 12-cube contains the desired point. The second bit specifies which of the 10-cubes is in question, and so on until a unique point has been determined.

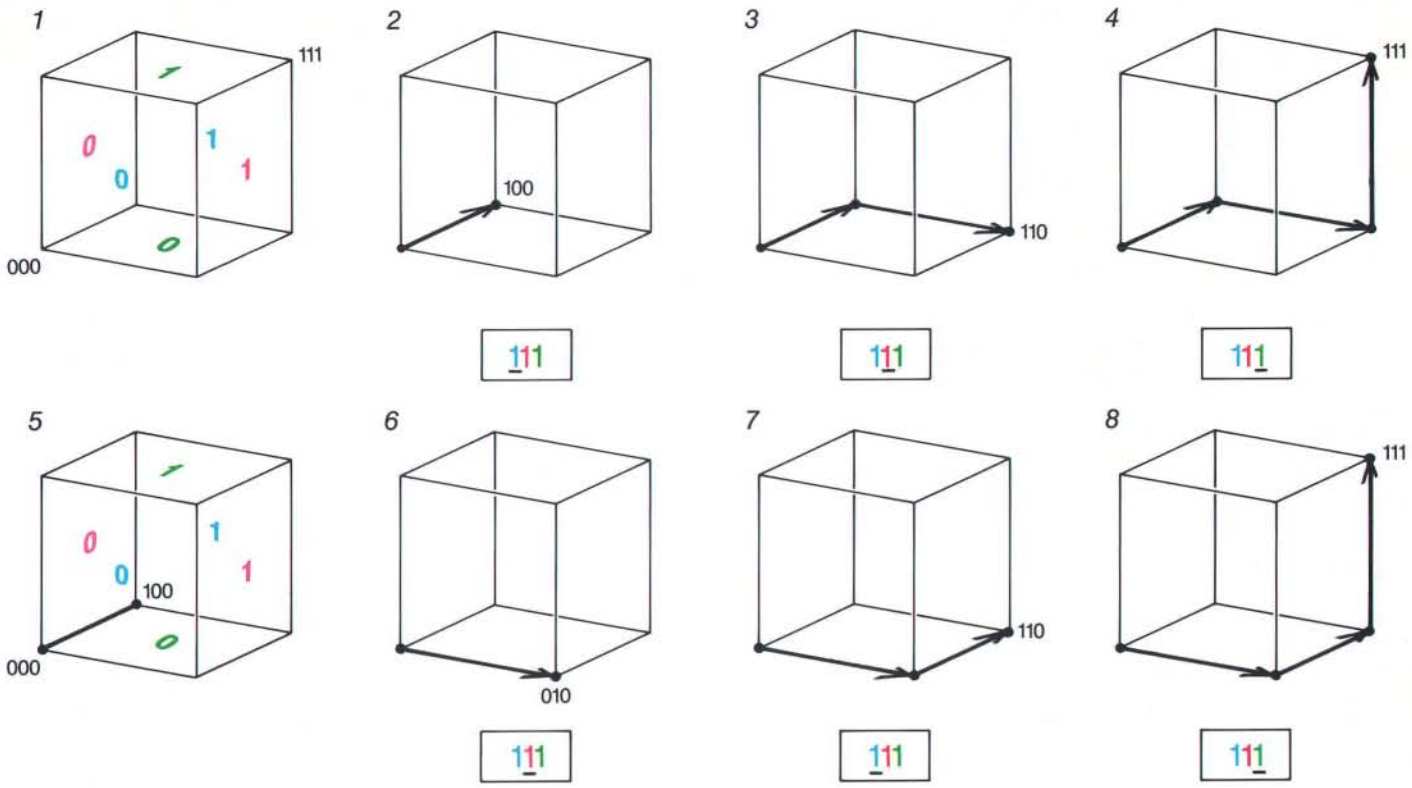


BOOLEAN *N*-CUBE provides the topology for the network that links the Connection Machine's processors. A Boolean *n*-cube is a generalized version of an ordinary cube. Such cubes can be constructed in many dimensions, each building on the next-lower dimension. A point can be considered a cube in zero dimensions, or a "zero-cube." Linking two points yields a "one-cube," or a line. Linking a pair of "two-cubes" (*squares*) yields the familiar three-dimensional cube. Two three-cubes can be joined at their vertexes (*corners*) to form a "four-cube." Repeating the process would yield a "12-cube" with 4,096 vertexes. The 4,096 chips of the Connection Machine are wired in the form of a 12-cube.

These binary addresses can be employed to route messages among the 4,096 chips in the Connection Machine. Each message in the system includes such an address. On receiving the message, the router examines the address one bit at a time, then forwards it to the next router along the way. That router in turn takes up the message, examines the address and forwards it. Thus, in no more than 12 steps, any message will find its way to the destination.

This communications network has several features that augment its speed and flexibility. One of the valuable properties of the *n*-cube arrangement is that there are many equally efficient routes of communication between any pair of processors. If one route is already occupied by a transmission in progress, the router is free to select an alternate route merely by processing the bits of the address in a different order [*see illustration on opposite page*].

Another type of flexibility is also inherent in the communications system. In some instances the communications network behaves more or less like a telephone exchange: it establishes a circuit between two processors so that they can communicate continuously and exclusively. In complicated cases, however, the messages may be so long and the system so crowded that the routers must behave more like post offices, storing packets of information that are later forwarded. Such decisions are made by the routers based on what wires are available when a transmission must be made.



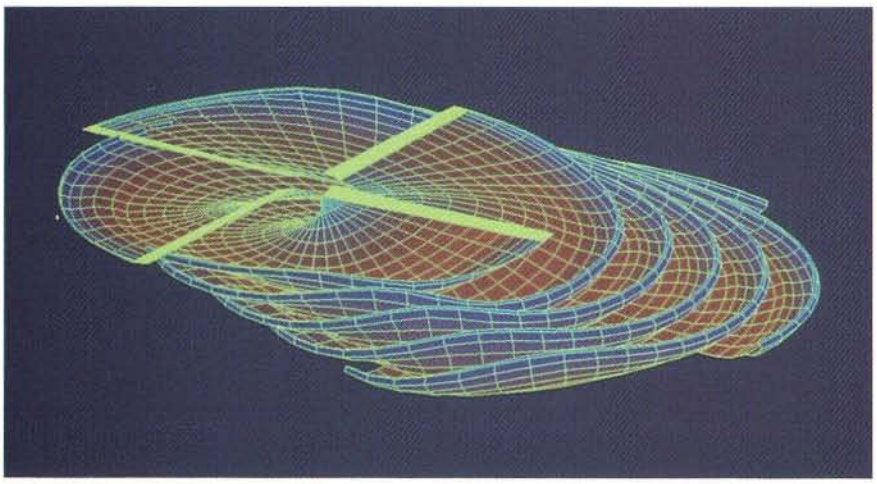
ALTERNATE ROUTES for communication between chips are provided by an *n*-cube. The illustration shows alternate routes in a three-cube, but the same principle applies to the 12-cube of the Connection Machine. Each vertex of the *n*-cube (where the chips lie) can be assigned a unique address as follows. A three-cube includes three pairs of planes. Each plane can be designated 0 or 1, and a vertex is then assigned a three-digit address according to which member of each pair of planes it is found in (*1*). Messages are forwarded by routing devices at each vertex, which read the address and process it one digit at a time. Here a message is sent from 000 to 111. The routing device begins by reading the first digit, and then it forwards the message to point 100 (*2*). There the second digit is read (*3*). At 110 the third digit is read, and the message is forwarded to its destination (*4*). When it comes time to send the message, however, the wire between 000 and 100 may be busy (*5*). In that case, the router simply reads the second digit of the address first, choosing an alternate route (*6*). Then the first and third digits are read (*7, 8*), and the message is delivered to the correct address.

These properties make it possible for the Connection Machine to establish many different patterns of communication, depending on the problem at hand. An important feature of the system is that such details are invisible to the user, who needs to know no more about Boolean *n*-cubes than the average user of the telephone needs to know about digital switching. (Indeed, future versions of the machine may have other wiring patterns with no effect on the algorithms that are employed.) The programmer interacts with the Connection Machine through a conventional computer, known as the host, which employs a standard operating system and programming language. The processors of the Connection Machine are connected with the host much as a conventional memory unit would be.

Indeed, in one sense the Connection Machine is the memory of the host. That relation makes possible a simple integration of parallel computing and existing software. Programs for the Connection Machine are surprisingly similar to conventional programs. The chief difference is that many operations normally carried out by repetitive loops are replaced by single operations corresponding to the simultaneous operation of many processors in the Connection Machine; the routing hardware automatically establishes the necessary communication paths.

It should be noted that nowhere in this system is exotic hardware to be found. In designing the Connection Machine, we chose well-tested technologies in order to achieve simplicity and reliability. The individual processors are relatively slow by the standards of today's fastest computers. The custom chip is built by methods similar to those for making personal computers and pocket calculators. Yet the assembled power of the 65,536 processors makes the machine very fast. For many applications, the machine can perform more than two billion operations per second; for the most favorable applications, the figure is more than 10 billion, or about 1,000 times as fast as a typical mainframe computer.

Putting the machine's speed in a slightly different context, one might consider floating-point operations, which provide a standard for computing power in number-intensive scientific applications. A floating-point operation is the multiplication or addition of two numbers expressed in scientific notation (such as 1.5×10^2). A



HELICOPTER ROTOR produces a complex airflow that can readily be simulated in parallel on the Connection Machine. Each of the machine's processors models the circulation within a certain layer of air (*small subdivisions of the image*). The circulation of the air in each section influences the air in each of the other sections. These interactions are computed in parallel. Details such as the wavy distortion at the bottom are important for predicting forces exerted on the helicopter blades. The simulation of the rotor was developed by T. Alan Egolf of the United Technologies Research Center and the author's colleague J. P. Massar.

typical supercomputer can carry out a few hundred million floating-point operations per second; the Connection Machine can average about 2,500 million on a typical problem.

To what uses has this considerable number-crunching capacity been put? As suggested above, some of the initial applications have involved manipulating and processing images. Others have exploited the parallelism inherent in certain physical processes. The engineering problem of calculating the flow of air over an airplane wing or a helicopter rotor provides an example of how the Connection Machine can mirror the parallelism of nature.

In nature the overall flow pattern emerges from the myriad interactions among air molecules, which bump into one another and into the surface of the wing as they rush along. The engineer (who is interested in the overall flow rather than the specific molecular interactions) uses a simplified, large-scale model consisting of a set of partial differential equations. Yet the equations themselves are set up in parallel: they treat changes in pressure in small volumes of air and sum their interactions to yield the overall flow. Because the equations are parallel, their solution is fast and efficient on the Connection Machine.

With a parallel computer, however, one can also move beyond the equations and come closer to the underlying physical reality. The large-scale behavior of a fluid is for the most part independent of the detailed physical properties of its individual particles. Moreover, the qualitative behavior of the fluid is not changed when the number of particles is greatly reduced. Therefore, it is possible to re-create accurately large-scale flows by examining collisions among a few tens of millions of simple, generalized particles.

Stephen Wolfram of the Center for Complex Systems Research at the University of Illinois at Urbana-Champaign and my colleague James Salem took advantage of this technique to model fluid flows over complex surfaces. Their simulation entailed only a few tens of millions of particles, and the particles were allowed to move in only six directions at integral velocities. Notwithstanding these limitations, the system is capable of accurately mimicking the flow of a fluid.

The simplest and most logical way to perform the fluid-flow computation would be to assign each particle its own processor. Yet a typical simulation includes about eight million "particles," and the Connection Machine, vast as it is, includes only some 65,000 processors. The solution to this programming difficulty and analogous ones is to program each processor to act as if it were a string of different processing units, each unit handling one particle at a time. The details of the arrangement are again invisible to the programmer, who simply specifies how many "virtual processors" are required. The hardware and software take care of the rest. Of course, if each processor must simulate 250 units in turn, the computation takes 250 times as long as it would with one actual processor per particle.

Many interesting applications of the Connection Machine do not involve numbers. My colleagues Brewster Kahle, Craig Stanfill and David Waltz exploited the computer's parallelism to retrieve documents from large collections of texts. The underlying principle of their system is that each processor can be programmed to compare one document in a large data base with a "search sample," a paragraph chosen for its relevance. Once the comparison has been made, the processors exchange information and rank the documents according to how well they match the search sample.

Comparing two pieces of prose to see how well they match is not a simple task. Merely counting the number of words that appear in both samples is useless, because the count is contaminated by words such as "the" and "as," which carry little content. Therefore, the document-retrieval system exploits a dictionary and some rules of grammar to extract from each sample the phrases that bear its content. Each processor is loaded with a different article compressed in this way, and the search sample is broadcast to all the processors in the network.

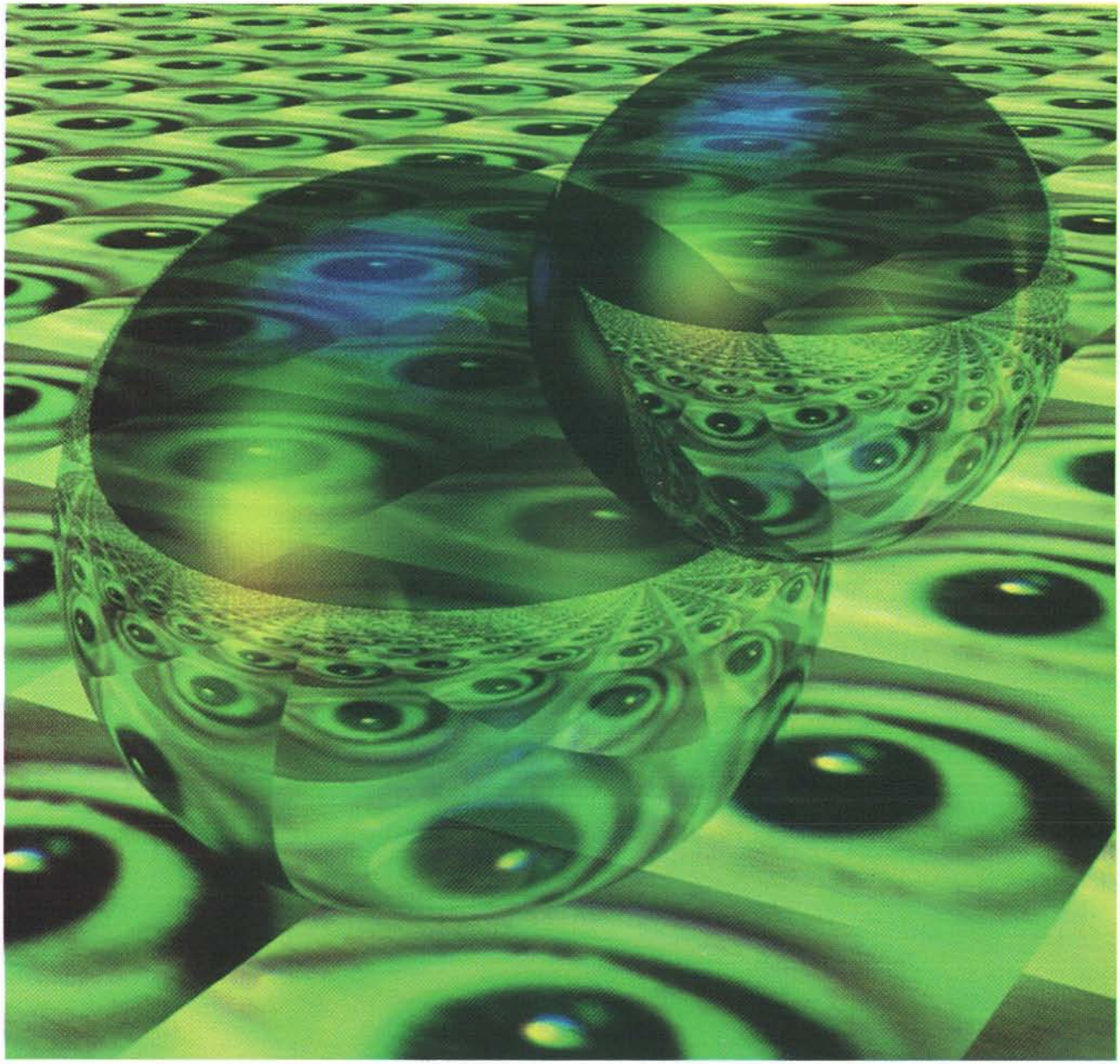
The process of comparison is relatively simple, and since 65,536 documents are checked at once, the entire data base can be examined almost instantaneously. Ranking the articles according to how well they match the search sample is a more difficult operation, because it requires a complex pattern of communication among the processors. Yet in parallel it can be performed in about 50 milliseconds. A few of the highest-ranked documents are then offered to the user of the system, who can choose a new search sample from among them. (Conversely, articles that are clearly irrelevant can be chosen as negative search samples.) Because all the comparisons are done at once, the entire collection of texts can be winnowed repeatedly in a short period, which ensures that all relevant articles will be found.

The document-retrieval program is able to function without anything that

approaches an understanding of the contents of the articles. Actually understanding those contents would require considerable background knowledge about the world, and that material has not been incorporated into the retrieval system. One exciting area of research involving the Connection Machine is the writing of programs that include such background knowledge and are able to mimic certain aspects of human reason.

Like the processing of two-dimensional images to form a three-dimensional world model, "commonsense" reasoning is carried out without apparent effort in the human brain. For example, any child can deduce (with appropriate affective response) that his mother's favorite vase will fall if it is dropped. The child is able to infer that the vase is more like a plate or a rock, which fall, than it is like a bird or a balloon, which do not. The inference can be made correctly in spite of apparently contradictory information, such as the fact that the vase may be spherical, as a balloon is, or that the child's mother may also own a bird.

Clearly, for human beings, the ease and accuracy of such inferences increases with the accumulation of knowledge about the world. The opposite is true for conventional computers. As the number of different



COMPUTER GRAPHICS is one of the fields in which parallel computers may be most fruitful. The illustration was made by the technique called ray tracing. In employing the technique, each Connection Machine processor is assigned to a different pixel (an abbreviation for "picture element"). The processors trace rays of light bouncing among imaginary objects; here the objects are glass balls and images of eyes. The paths of the rays determine the final color of each pixel. Karl Sims of the Massachusetts Institute of Technology Media Laboratory generated this image.

concepts increases, the number of possible relations among them increases even more quickly. Because a sequential computer can examine these relations only one at a time, its pace slows dramatically as the quantity of background data grows. Indeed, the slowness of conventional computers in commonsense reasoning was one of the stimuli responsible for the design of the Connection Machine.

In the late 1970s, as a graduate student at the Massachusetts Institute of Technology, I became interested in how commonsense reasoning might be simulated by computers. It seemed to me that one way out of the morass sequential computers found themselves in when asked to make simple, everyday deductions was to build a machine that could examine possible connections among concepts more than one at a time. (In that conclusion I was inspired by the work of Marvin L. Minsky of M.I.T. and Scott E. Fahlman of Carnegie-Mellon University.) That was in 1978. By 1985 the idea had moved to the stage of an actual prototype with the aid of a grant from the U.S. Defense Advanced Research Projects Agency, which offered to buy the first machine. By then I had left M.I.T. and helped found the company—Thinking Machines Corporation—that builds and markets the Connection Machine.

Now that the Connection Machine is a physical reality, investigators of artificial intelligence are making use of it to solve commonsense reasoning problems. The germ of this approach is to assign one fundamental concept to each processor. The connections among processors can then be exploited to represent multiple relations among simple concepts. In the simple example given above, one processor may represent the concept "Vase," another "Mother" and a third "Likes." The connections among these three processors would embody the knowledge that "Mother likes her vase." Other connections might represent the vase's shape, composition and history. When it comes time to decide what the outcome would be if the vase were dropped, the relevant connections can be searched in parallel.

Since there are now about a dozen Connection Machines in operation, there will undoubtedly soon be many new programs for the machine. It is likely that many of them will be in the four general areas touched on above: image processing, simulation of physical processes, searching of data bases, and artificial intelligence. One of the greatest challenges in learning to use the Connection Machine lies in beginning to think in parallel terms. Programmers have considerable accumulated experience in programming for sequential machines, and such programming has by now become almost second nature. Learning to write programs for parallel machines requires thinking in ways that are quite different from those demanded by sequential computers.

That challenge will be greater for the Connection Machine than it will be for some other types of parallel machines. It should not be assumed that the Connection Machine is the only representative of its genre. Indeed, many different parallel designs are now in various stages of realization. To generalize greatly, these designs fall into two broad classes: "coarse-grained" and "fine-grained." Coarse-grained machines link relatively few processors, each with a relatively large amount of computational power; fine-grained machines link a large number of weak processors.

These two classes of parallel computers form a spectrum. At one end is the conventional sequential computer, which has the minimum number of processors: one. At the other end of the spectrum are designs such as that of the Connection Machine, which include a very large number of small processors. Although some highly qualified investigators and companies are pursuing the coarse-grained approach, I think it is the fine-grained design that will ultimately prove the most fruitful. Yet it is also the one that is the most foreign to our preconceptions about computer programming.

In writing a program for a coarse-grained machine, one can adhere to concepts much like those used for programming sequential computers. The problems in this work arise in attempting to coordinate the programs. In writing a program for the Connection Machine, however, one is faced with an entirely different realm of problems and possibilities. Exploiting the full potential of the machine will require a new way of thinking about computation, which we as programmers have just begun to learn. That learning process will undoubtedly be both rewarding and challenging.

Some of its rewards may come from the fact that the Connection Machine can be expanded to encompass much more computational power without any fundamental changes in design. Most of the applications envisioned for the machine could profitably exploit a computer much larger than current versions of the machine. Hence, the computer has been designed to allow a significant increase in the number of processors. The Connection Machine can be expanded simply by adding processors, memory and communication devices to an existing machine.

As an extreme example of scaling up, imagine a parallel computer with one billion processors. Such a machine might well incorporate some features of the Connection Machine, although there would undoubtedly be many new problems to solve. If built with current technology, a billion-processor machine would be as large as a building and cost 20 times as much as today's largest computers. It could, however, execute some 100 million million (10^{14}) instructions per second, which is several orders of magnitude greater than the computational power of all existing supercomputers combined.

There are technical problems inherent in building such a computational engine, but they are solvable. The real problems are those of the imagination: conceiving how such power would be used. Some engineering problems, including extrapolations of examples mentioned above, might benefit from such capacity, but they are in a sense trivial. The applications worthy of a billion-processor machine are those that entail a radical change in the way we think about computation.

A parallel computer with a billion processors might provide the basis for a computational utility analogous to existing gas and electric utilities. Just as a coal-fired plant generates electricity that is transmitted to individual appliances, a huge parallel computer could provide computational power to a city's worth of robots and workstations. The design of the parallel machine would enable many users to draw on portions of the total computing capacity for small problems, whereas the total capacity could be applied to large ones. Such a vision is somewhat utopian now, but it is by no means impracticable, which suggests the depth of the changes that parallel computing may ultimately bring.

FURTHER READING

THE CONNECTION MACHINE. W. Daniel Hillis. The MIT Press, 1985.

MASSIVELY PARALLEL COMPUTERS: THE CONNECTION MACHINE AND NON-VON. Richard P. Gabriel in *Science*, Vol. 231, No. 4741, pages 975–978; February 28, 1986.

SPECIAL ISSUE ON PARALLELISM. *Communications of the ACM*, Vol. 29, No. 12; December 1986.

SCIENTIFIC AMERICAN®

The cover shows NGC 2997, one of the millions of galaxies in the observable universe. Photograph from the Anglo-Australian Telescope Board, photography by David Malin.

SCIENCE IN THE 20TH CENTURY

SCIENTIFIC AMERICAN *Science in the 20th Century* is published by the staff of SCIENTIFIC AMERICAN, with project management by:

EDITOR: Jonathan Piel

PROJECT EDITOR: James T. Rogers

ART: John McLaughlin, *Project Art Director;* Joan Starwood, *Associate Art Director*

COPY: Maria-Christina Keller, *Copy Chief;* Nancy L. Freireich; Jonathan Goodman; Daniel C. Schlenoff

PRODUCTION: Richard Sasso, *Vice President Production and Distribution; Managers:* Carol Albert, *Prepress;* Tanya DeSilva, *Projects;* Carol Eisler, *Manufacturing and Distribution;* Carol Hansen, *Composition;* Madelyn Keyes, *Systems;* Leo J. Petruzzi, *Manufacturing and Makeup;* Jo Marie Fecci and Sue Griffin-Porritt, *Typesetting;* William Sherman, *Production Coordinator*

CIRCULATION: Lorraine Leib Terlecki, *Circulation Director;* Cary Zel, *Circulation Manager;* Rosa Davis, *Fulfillment Manager;* Katherine Robold, *Assistant Business Manager*

ADVERTISING: Robert F. Gregory, *Advertising Director;* Meryle Lowenthal, *Advertising Manager/* SCIENTIFIC AMERICAN *Science in the 20th Century;* Laura Salant, *Sales Services Director*

ADMINISTRATION: John J. Moeling, Jr., *Publisher;* Marie D'Alessandro, *Business Manager*

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue
New York, N.Y. 10017
(212) 754-0550

PRESIDENT AND CHIEF EXECUTIVE OFFICER: Claus-Gerhard Firchow

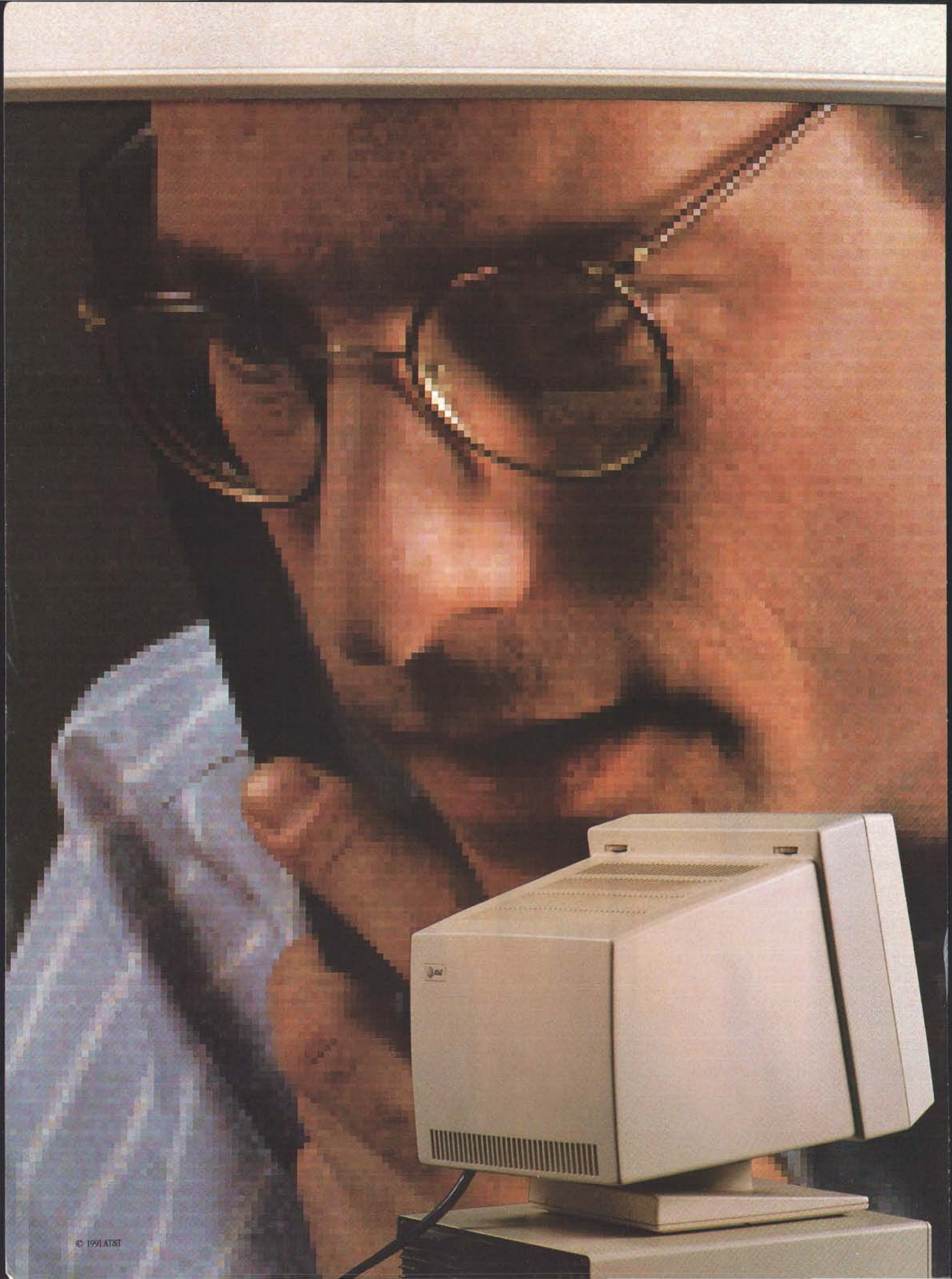
CORPORATE OFFICERS: *Executive Vice President and Chief Financial Officer,* R. Vincent Barger; *Senior Vice President:* Linda Chaput; *Vice Presidents:* Jonathan Piel, John J. Moeling, Jr.

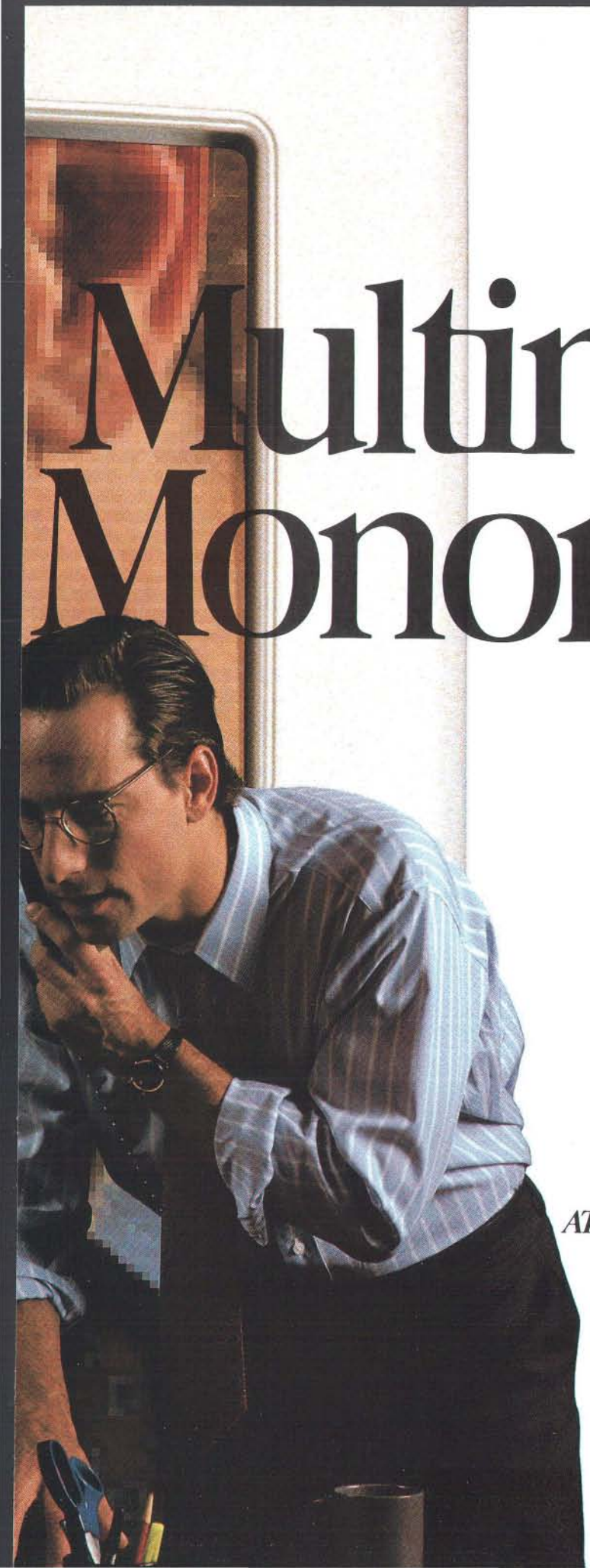
CHAIRMAN OF THE BOARD: Dr. Pierre Gerckens

CHAIRMAN EMERITUS: Gerard Piel

THE ILLUSTRATIONS

Page	Source
5	California Institute of Technology and Carnegie Institution of Washington
6	Jason Küffer
7	Wolfram Research, Inc.
8	Frederic M. Richards and Paul E. Vogt, Yale University
9	Peter J. Mouginis-Mark, University of Hawaii
10	Laurie Grace
17	National Bureau of Standards (*top*), Radio Corporation of America (*bottom*)
18–21	Irving Geis
23	Argonne National Laboratory (*top*), Ilil Arbel (*bottom*)
24–31	Gabor Kiss
33–39	Ian Worpole
43	Ben Shahn
49–61	Ian Worpole
63–69	Andrew Christie
73–78	Joe Lertola
82–83	Sara Love
84	Paul Weller
85	H. R. Wilson (*left*), R. E. Franklin (*right*)
86–88	Sara Love
93–94	Robert Langridge
95	Nelson L. Max
96	Robert Langridge (*top*), Arthur J. Olson (*bottom*)
97	Richard J. Feldmann
98–101	Ilil Arbel
105	George V. Kelvin
106–110	Michael Goodman
111	Cetus Corporation
115	Royal Canadian Air Force
116	Andy Christie
117	G. S. Johnson, H. M. Geological Survey (*top*), Royal Canadian Geographical Society (*bottom*)
118–122	Andy Christie
123	Royal Air Force
124–127	Andy Christie
128	Laurie Grace
132–133	Ian Worpole
134–135	Andrew Tomko
136	William F. Haxby, Lamont-Doherty Geological Observatory of Columbia University
137	Ian Worpole
138	Andrew Tomko
142–143	Stephen P. Miller, Daniel S. Scheirer, Charles M. Weiland, Suzanne M. Carbotte and Laura J. Perram, University of California, Santa Barbara, and Stacey A. Tighe, University of Rhode Island
144	Ian Worpole
145	Nancy R. Grindlay, University of Rhode Island, and Suzanne M. Carbotte
146	Ian Worpole
147	Laura J. Perram
148	Ian Worpole (*top*), Laura J. Perram, Daniel S. Scheirer and Marie-Helene Cormier, University of California, Santa Barbara
149	Ian Worpole
150	James Lewicki
151	Bell Telephone Laboratories
152–153	James Lewicki
155	Steve Allen, Fairchild Semiconductor
156	Motorola Semiconductor Products, Inc.
157	Ben Rose
158	Ben Rose (*top left*), Steve Allen (*top right*), Texas Instruments, Inc. (*bottom*)
159	Steve Allen (*top left*), Dan Todd (*top right*), Texas Instruments, Inc. (*bottom*)
160–162	Andy Christie
164	Phillip A. Harrington
166–172	Gabor Kiss
175–176	Stephen Grohe
177–179	Gabor Kiss





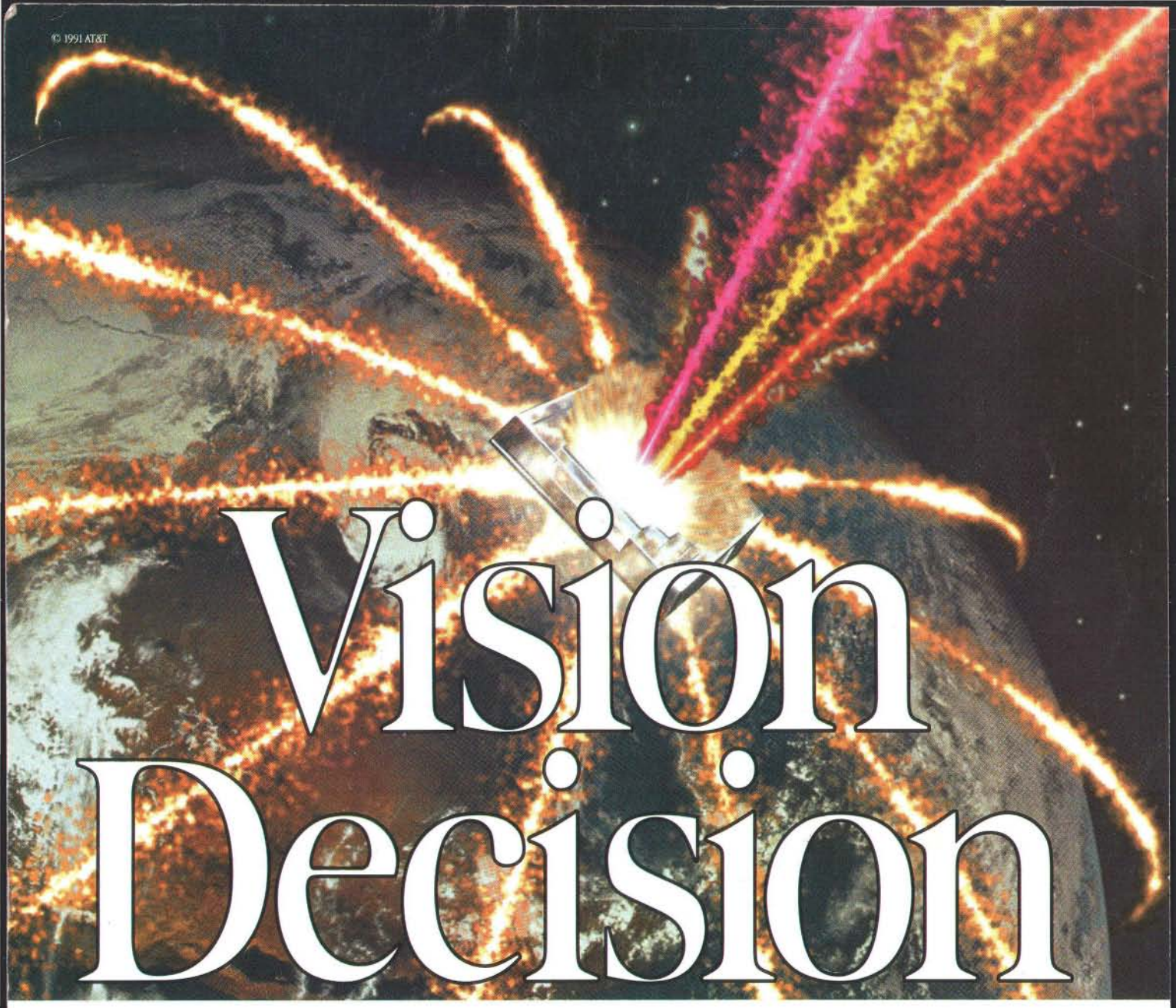
Multimedia Monomania

Or, Why We're Absolutely Positively Obsessed With Helping Computers Listen, See, and Talk.

One thing is clear about the future of technology. There will be a merger of voice, data, and image communications. We'll "talk" to our computers rather than "key" them. "Listen," rather than "read" them. High speed compression and decompression, digital signal processing, word spotting, ISDN, and other technologies will make this multimedia world possible. And what these unique technologies have in common is that AT&T Network Systems and Bell Laboratories are in the forefront of all of them. AT&T is obsessed with researching. Developing. Integrating. Implementing. So when multimedia is finally "here," you'll find it at AT&T and your local phone company first.

AT&T and Your Local Phone Company Technologies For The Real World.

AT&T
Network Systems



Or, How AT&T's Global Vision Helps Your Company Make The Right Decision.

Inside this special edition of Scientific American, you'll discover how AT&T's vision of Universal Information Services (UIS) can make your business run better. UIS is a totally connected world. Where voice, data, and image processing become one and the same. It's a unique combination of AT&T Bell Laboratories technologies like broadband and photonics. Technologies that let you send any kind of information. Anywhere. Anytime. No ifs. Ands. Or buts. So your far flung company can work as an integrated, synergistic whole. UIS enables you to get just where you're going, without having to throw out what you already have. Or will have. That's why AT&T's vision is the right decision.

AT&T And Your Local Phone Company
Technologies For The Real World.

AT&T
Network Systems